UNITED STATES DEPARTMENT OF THE INTERIOR
RAY LYMAN WILBUR, Secretary
OFFICE OF EDUCATION
WILLIAM JOHN COOPER, Commissioner

RESEARCH IN HIGHER EDUCATION

Papers prepared for the First Regional Conference on Higher Education held under the joint auspices of the United States Office of Education and the University of Oregon at Eugene, Oreg., April 14, 15, and 16, 1931

BULLETIN, 1931, No. 12

UNITED STATES
GOVERNMENT PRINTING OFFICE
WASHINGTON : 1932

FOR SALE BY THE SUPERINTENDENT OF DOCUMENTS, WASHINGTON, D. C.

CONTENTS

LETTER OF TRANSMITTAL

DEPARTMENT OF THE INTERIOR,
OFFICE OF EDUCATION,
Washington, D. C., May, 1932.

SIR: About two years ago it occurred to me to attempt a conference on research in higher education. Accordingly, I discussed the matter with Dr. Arnold Bennett Hall, President of the University of Oregon and we decided to hold a preliminary conference on this subject in April, 1931. Men were gathered from the entire Northwest for a 2-day conference on the subject. Some of the papers were of considerable value in opening up the subject. These papers were brought together and disposed of in some four ways. Here are the ones which we consider worthy of printing as beginning our work in research in higher education. I recommend that these be printed as a bulletin of the Office of Education.

Respectfully submitted.

WM. JOHN COOPER,
Commissioner.

The SECRETARY OF THE INTERIOR.

FOREWORD

The application of scientific method to the study of college problems in the fields of curricula, methods, administration, and student personnel is increasing rapidly. The experimental method—the setting up of alternative procedures in such a way that the factor under study may be isolated for measurement—appeals more to college faculty members than other methods of investigation. It is the method to which many of them are accustomed in their academic fields. It is the method which allows for greatest objectivity, and hence carries greatest weight.

In higher education, there are many long-established practices and well-settled convictions. These may be molested only in the light of definitely authenticated facts. Faculty discussions tend to be fruitless because of the absence of such authentication. If the present general unrest in the colleges is to result in wise changes, there must be widespread experimentation with the problems involved.

Recognizing the fundamental importance of this experimentation, the Division of Colleges and Professional Schools of the United States Office of Education hopes to shape its program so as to be of as great assistance as possible in stimulating experimentation in the universities and colleges. The conference in the Pacific Northwest at which the papers published in this bulletin were read, was the first of what it is hoped may be a series of regional conferences where results of experimentation may be reported and discussed. The present bulletin is the first of what it is hoped may be a series of bulletins to be published by the United States Office of Education, making available the results of experimental studies in higher education.

The facilities of the library in the Office of Education and the services of the staff in the Division of Colleges and Professional Schools will be available to aid in forwarding this program of experimentation.

FRED J. KELLY,
Chief, Division of Colleges and Professional Schools.

RESEARCH IN HIGHER EDUCATION

INTRODUCTION

HENRY DAVIDSON SHELDON
Dean of the School of Education, University of Oregon

The papers here published were read in connection with the program of the First Conference on Higher Education for the Pacific Northwest, held at Eugene, Oreg., April 14 to 16, 1931, under the joint auspices of the United States Office of Education and the University of Oregon. Those prepared by members of the University of Oregon staff, together with an earlier collection by the university under the title of "Controlled Experimentation in the Study of Methods of College Teaching," constitute the permanent results of a 5-year experimental program for the Improvement of College Teaching in the University of Oregon. The fundamental idea back of this program has been not so much the production of permanent contributions to the scientific literature of this field, although it is hoped that this may be a by-product, but rather the stimulating of interest on the part of the teaching staff of the university in a more effective point of approach in handling the teaching situation. Consequently, at this time there may be a certain value in endeavoring to summarize the results, as far as they have been achieved, from this point of view.

The first problem attacked in this program at the University of Oregon was the problem of securing greater initiative and more alert intelligence on the part of the students in college courses. The criticism heard frequently that the prevailing modes of instruction in college, whether based on lectures, textbooks, or library reading, were largely passive and that the student is engaged in absorption rather than discovery, was felt to have much weight. The protests against this condition in various fields of education are well known. For the preschool age, Madam Montessori; for the elementary and junior high school, the problem-project method of approach; in science instruction, the laboratory method, have all been designed to improve this situation. In the field of higher and professional education, the law schools of the country, through the employment of the case method of instruction, seem to have made the most distinct progress in this field. In the University of Oregon, several departments undertook to organize experiments which should endeavor to show the value of a set-up involving more initiative. The departments in question

were psychology, economics, sociology, and education. Some of the results of these experiments have been printed in the earlier volume. The experimentation has been continued but without material addition to the results there published. Apparently the case problem-project method secures more response and better results from the able students without materially lessening the achievement of those belonging to the inferior and mediocre groups. More than this we can not say until it has been possible to work out through several years of special appliance of a variety of techniques on a college level for this new type of work, something which the law schools and schools of business administration have already done.

In a period when the limitation of numbers has become, because of the expense of higher education, a very pertinent problem to college administrations, the question of means of selecting students naturally comes to the forefront and it becomes important to discover to what extent it is possible to determine in advance the scholastic future of candidates. The University of Oregon has worked on this problem from two points of view. The establishment of intelligence tests for all entering students during the last five years has furnished material for the study of the value of these tests as a means of prediction. Dr. Howard Taylor, who has had charge of this work, has organized and interpreted this material in one of the papers in the present collection. Considerable attention has also been given to the second aspect—that of special aptitude tests. Here two tests, one in accounting by Professor Stillman and one in English by Mr. Shumaker, have been developed to a point of usefulness beyond previous contributions in this field. Beginnings have been made in two other fields, journalism and teaching. Experience has shown that success here depends on a very large amount of empirical experimentation and can only be secured at great expense of time and energy through several years.

The reliability of grades and the improvement of current examination systems constitute another province for investigation. While it has been the purpose of the committee in charge to encourage a much larger use of objective examinations than has hitherto been the case; it was also felt that much might be done in the improvement of the so-called essay type of examinations. The cooperation of the university administration in asking all departments to file copies of each examination has placed a large reservoir of material at the disposition of the committee. Two papers in the present collection by Clifford Constance and Ralph Leighton represent the contribution here.

The university has also cooperated in the movement, general throughout the country, to assist freshmen and other beginning students in study procedures through reading tests, schemes for the organization of time, use of library, etc. The number of students

exposed to this procedure is not large but the results which are given in a paper by Messrs. Tuttle and Davis are reassuring. Other aspects of this problem are dealt with by the papers of Professor Parr, of Oregon State College, and Professor Jordan, of State Teachers College, Ellensburg, Wash. To secure further results and values here, there should be a much larger number of cases which should be studied individually and segregated by types. The fact that many colleges and universities are laboring in this field suggests the possibilities of cooperative institutional research.

In addition to the lines of endeavor already mentioned, there have been certain miscellaneous experiments such as the one in the field of English history on the value of formal quizzes, published in the early collection by Dr. Donald Barnes. These studies are being continued and certainly point to the necessity of certain changes in procedure.

From the point of view of understanding present conditions in the University of Oregon, there has been begun a series of investigations of a collective and local sort not represented in the present volume, because of the late hour of completion. In an institution of large size the actual results of the teaching procedures are difficult to determine and student and faculty gossip concerning them is by no means trustworthy. Consequently, the committee in charge of investigating college teaching has organized a number of studies to determine the results of certain devices in teaching. The procedures taken up in this way have been term papers, investigated by a special committee of which Prof. L. L. Lewis, of the English department, was chairman, and the actual amount of time spent in assigned readings in book reserve by Prof. Virgil Earl. Certain departments of the university depend on this method exclusively for students' study. The value of segregation according to ability in certain sections has also been studied and reported on.

The investigations of college teaching at the University of Oregon have been entirely the work of a large faculty committee of 12 members representing many of the important schools and departments of the university. At first this committee was made up largely of department heads and deans of departments. It was discovered, however, in course of time that, after the general policy had been outlined, more fruitful results could be obtained by a membership made up in the main of younger men in the departments who have the time and leisure to undertake studies themselves. Consequently the committee is more and more becoming a clearing house for the discussion of technique along the lines represented.

The university administration has assisted by wholeheartedly supporting the program of the committee and also by furnishing small sums of money which are necessary for mimeographing certain mate-

rial, printing tests, and for the mechanical and statistical manipulation of the results of the experiment. The institution has also supplied technical statistical advisers for consultation purposes. These two forms of practical aid are undoubtedly necessary if a program of this sort is to be carried out effectively.

In diffusing the results of the investigations, of conferences, and visits to other institutions, the committee has adopted the plan of holding special meetings of the faculty known as "colloquia" where the results are presented and discussed. Some of the meetings have been largely attended and have constituted a valuable stimulus. As the program has proceeded, the necessity for these formal meetings has largely disappeared as it is found that the spontaneous interest of each group in a particular experiment creates a more genuine reaction than any formal meeting could secure.

The committee has also endeavored to post the faculty through the publication of mimeographed bibliographies, special book reviews, and through the establishment of a special shelf in the library with duplicate copies of important books and periodicals on college teaching. The personnel office under the direction of Dr. Howard Taylor has also assisted by sending out mimeographed studies of certain local problems based on material collected.

The results of a campaign of this sort are somewhat difficult to estimate. Undoubtedly a more analytic attitude toward the entire subject of college teaching has resulted with a majority of the faculty. Two-thirds of the departments and schools of the university have at one time or another participated in certain experimental activities and at least one-third have carried on somewhat systematic experimental programs. This indirect method of approach, depending on the voluntary activity of members of the faculty, is undoubtedly slower in securing results than certain other more direct campaigns involving the teaching staff in compulsory systematic activities but it is felt that it avoids a large amount of antagonism and probably in the long run is more pervasive in changing the attitude of the instructing staff.

GROUP I.—INSTRUMENTS OF MEASUREMENT IN THE FIELD OF COLLEGE INSTRUCTION

HOW RELIABLE ARE COLLEGE MARKS?[1]

HOWARD R. TAYLOR[2] and CLIFFORD L. CONSTANCE[3]

If college marks were merely rewards for the more or less faithful performance of academic tasks, they would probably not merit scientific study. Since, in general, marks have become increasingly important as a basis for administrative and educational procedures—in American colleges at any rate—some determination of the credence which they deserve is likewise increasingly important. Faculties and administrators usually put their requirements of students in terms of marks. Students are dismissed, prevented from transferring to other institutions, ruled ineligible for athletic competition, awarded fellowships, accorded privileges such as working for honors, and so on—all on the basis of marks. Again, studies of learning suggest that a more less objective knowledge of results, such as marks attempt to provide for college instruction, is a positive factor in improvement. The selective function of colleges—a seldom recognized but perhaps major service of such institutions—is primarily dependent on the marking of students. Moreover, although our whole social organization is based upon individual differences in general capacity and special aptitude, school marks provide almost the only organized attempt to give individuals the information about themselves in comparison with others which is indispensable if ambitions are to be brought in line with actualities. Evidence that college marks perform such a selective guidance function with at least partial validity is plentiful. President Gifford (3)[4] of the Bell Telephone Co. has

[1] The writer of this paper is Mr. Taylor, but the study was made by Mr. Constance under the writer's direction as a master's thesis in psychology at the university.

[2] Howard R. Taylor, director, Personnel Research Bureau, University of Oregon. A. B., Pacific University, 1914; A. M., Stanford University, 1923, Ph. D., 1927. Publications: "The Need for Personnel Research in a University," *School and Society*, 26:673, Nov. 19, 1927; with F. F. Powers, "Bible Study and Character," *Pedagogical Seminar and Journal of Genetic Psychology*, 294–302, June, 1928; "The Influence of the Teacher on Relative Class Standing in Arithmetic Fundamentals and Reading Comprehension," *Twenty-seventh Yearbook of the National Society for the Study of Education*, Part II, chapter 5, 97–110, 1928; "An Experiment with Independent Study," *Controlled Experimentation in the Study of Methods of College Teaching*, University of Oregon Publication, Education Series, 1:7:300-312, February, 1929; "Teacher Influence on Class Achievement: A Study of the Relationship of Estimated Teaching Ability to Pupil Achievement in Reading and Arithmetic," *Genetic Psychology Monographs*, 7:2, February, 1930.

[3] Clifford L. Constance, assistant registrar, University of Oregon. B. A., University of Oregon, 1925, M. A., 1929. Publication: "Greeks of the Campus," *School and Society*, 30:409–414, 1929; "Personality Ratings Given High-School Graduates by Principals and Teachers," *School Review*, 39:683–688, 1931.

[4] Numbers in parentheses refer to "Bibliography," p. 14.

shown why his organization uses marks made in college work as a basis for employment—high marks tend to indicate the abler men. In college the guidance of students by marks into fields of study where their interests and ambitions harmonize with their abilities occurs continuously on both the conscious and unconscious level.

Hence, it would seem that these varied uses of college marks are necessary and reasonable enough if only the marks are dependable. In fact, most of the objections to marking students disappear in proportion to the reliability and validity of such judgments. Even the artificiality of marks is unimportant if only they measure accurately what they are supposed to measure. While the validation of college marks, i. e., the social and individual significance of excellence in the various fields of college instruction, is perhaps even more important than the matter of reliability, this study will be confined to the latter simpler but still very important issue.

When a surveyor measures off a city lot, he assumes reasonably enough (*a*) that his tape is equivalent to any other good tape within negligible limits; (*b*) that his errors in reading it are negligible, or at least that they can be made so by repeated measurements; (*c*) that the area and shape of the plot will remain constant year after year, or at least practically so. When psychologists attempt to apply scientific techniques to such measures as college marks, the extent to which analogous sources of error are negligible must be empirically determined.

When an instructor appraises student achievement, (*a*) it is seldom or never true that he uses exactly the same standards as other equally competent instructors. Nor can he apply the same examination again and again in the way a surveyor uses a steel tape. Hence, various scholastic tape measures are fearfully and wonderfully different one from another. (*b*) Seldom will two instructors when evaluating the same sample of student performance agree as to its merit. Even two successive appraisals of the same paper by the same instructor are likely to differ. Thus instructors are also unable to read very precisely such scholastic scales as they do have. Even in supposedly objective subjects, such as mathematics and science, disagreements in appraising performance are large, as studies by Starch, Ruch, Wood, and others have repeatedly demonstrated (7, 9, 11). The remedy as in all scientific measurement is to make repeated measurements so that successive errors of a chance sort will tend to cancel. Thus the sum of many relatively crude measurements may be decidedly better than the separate estimates of which it is composed. (*c*) Furthermore, the abilities and especially the performances of students are not the same yesterday, to-day, and forever. Students are affected by health, happiness, and countless other things so that a measure of a student's true achievement would necessitate many evaluations under

all reasonable conditions over a fairly long period of time. The average of such a series of measures might be considered the student's true ability. Especially during instruction with fluctuating effort, to which college students are prone, it is reasonable to expect considerable change both absolute and relative in student achievement from week to week and year to year.

Fortunately it is the relative size only of errors which has significance. An error of an ounce or two—even a few pounds—is usually negligible in weighing human adults because the differences between individuals are, in general, much greater than an error of this degree. If one were weighing mosquitoes, finer scales would be necessary, because the heaviest mosquito weighs only a small fraction of an ounce more than the lightest. Since in the case of college marks (*a*) inequalities in standards and errors in sampling scholastic performance, (*b*) errors in evaluating these samples of student performance, and (*c*) fluctuations in the performance of individual students are all inherent in the situations where the marks are to be used, it is the unreliability of marks as a result of all these sources of error that we wish to determine.

Now, if two similar measurements of the scholastic achievement of each student are made independently at reasonable intervals, and if each student tends to hold about the same rank in comparison with others, we may conclude that the errors from all these sources are small in comparison with the total range of individual differences in such achievement. Since the correlation coefficient expresses very compactly the extent to which one measurement of a group of individuals will predict another in linear fashion, it is customary to consider the correlation of two similar measurements made independently under representative conditions an index of the accuracy of the measurements.

In studying the reliability of college marks the practice has been to correlate successive quarters, semesters, or years of college work. But students often do not take closely similar programs of work, even in successive quarters and semesters, and certainly not in successive college years. Thus the kind of performance sampled when the excellence of work in two different terms is correlated, theoretically at least, is not sufficiently similar to indicate the real accuracy of the measurements. In so far as samples of dissimilar scholastic performance are involved, such comparisons give reliability coefficients which are too low. On the other hand, instructors are seldom able to make completely independent estimates of the merit of performance in successive terms of the course and students are likely to be rated twice on much the same basis when scholastic performance for two such terms is correlated. Then the comparisons are likely to give

spuriously high reliability coefficients because of correlated errors in the two series of estimates.

Finally, we can hardly expect the distractions of college life and the accompanying individual fluctuations in interest and effort to affect all students alike. Football affects the gridiron warriors and their partisans most in the fall quarter. For the basket ball men and their following whatever effects there are come chiefly in the winter term, while canoeists and poets are probably most susceptible to "spring fever" at that time of year. Thus correlations between quarters and semesters of college work furnish reliability coefficients which are too low because the fluctuations in individual performance for the periods compared are not typical.

In our study we have attempted to minimize these three (in part counteracting) sources of unrepresentative error in the assignment of college marks. We began with the records of 418 men and 403 women who entered as freshmen and completed the fall term of 1925–26 at the University of Oregon. Of these, 184 men and 212 women remained for two consecutive years, 1925–26 and 1926–27. Their scholastic records in these six quarters of lower division work furnished the basic data for the study. Thus we had six measurements—marks evaluating the scholastic work of six consecutive quarters—for each individual. These marks were then segregated by departments as well as by quarters for each student. We then took marks earned in the fall of 1925, spring of 1926, and winter of 1927 (the odd quarters of our six) and compared them with marks earned in the winter of 1926, fall of 1926, and spring of 1927 (the even quarters). We then paired these marks by departments, so that in each set of measures the kind of scholastic performance rated was matched in the other set, giving us two representative samples of the scholastic performance of each student for each department in which he took courses. Thus the kinds of work attempted are reasonably similar. The estimates have been made by different instructors, or at least separated by a 3-month interval of time, so as to favor independence in the judgments. Work done in similar quarters of different years has been utilized to make fluctuations in individual performance comparable. Wherever a student had two terms of work in a given department both of which happened to fall in the same set of terms, i. e., both in the odd or both in the even set, they of course could not be used. This occurred so seldom that only 276 out of more than 11,000 separate grades had to be discarded in the study.

Marks at the University of Oregon [5] are on a 6-step scale, as follows: I, unusual excellence; II, high quality; III, satisfactory; IV, fair; V, passing; VI or F, failure. If marks III and IV be combined as average, the scale is practically the same as the more widely used

[5] Such was the grading system at the time of this study. It has since been changed to A, B, C, D, and F.

5-step scale. The university catalogue states the expected proportions of the average class to which these various grades will be assigned as I and II, 20–25 per cent; III and IV, 55–65 per cent; V, 15–20 per cent; VI or F, not stated.

The actual distribution of grades at Oregon is skewed a little to the upper end of the scale, a fact which might be explained on the hypothesis that Oregon has a preponderance of above average students, but is probably an expression of the leniency of most professors in giving more than the expected proportion of students the benefit of scholastic doubts. The percentages of each grade assigned throughout the university are quite constant from year to year, and at the time of this study were I = 9 per cent, II = 23 per cent, III = 35 per cent, IV = 19 per cent, V = 9 per cent, and VI or F = 5 per cent. Since studies of college marks for large numbers of students over a considerable period of time at many representative institutions have empirically demonstrated a uniform tendency to approximate the normal curve in such distributions, we have determined the numerical weight to be assigned each mark by determining the mean deviation in a unit normal distribution of each portion corresponding to the actual percentages of each grade assigned at the University of Oregon (4). Of course, departments differ considerably in the percentages of students to which each mark is assigned, and there are selective factors which in some cases offset and in others aggravate the inaccuracies suggested by these differences. But since marks have been paired by departments, such errors will be practically constant for each individual in both sets of measures, and hence should not detract from the determination of their relative reliability. Moreover, our findings in general agree with those of Spence (8) when he says that the improvement produced by correcting for the variations due to different percentages of grades assigned by different instructors is not large, and again that while there are significant differences in the intellectual level of various classes, correction of grades for such differences at present makes very little difference in the final composite for any student.

The standard deviation values for each category of the grading scale according to the percentages of each grade actually assigned came out I = +1.82, II = +0.86, III = +0.04, IV = −0.69, V = −1.27, VI or F = −2.40. It can be shown mathematically that the relative value of these weights remains the same when the constant 2.4 is added to each so as to avoid the use of negative numbers in computation.[6] The weights thus computed would be I = 4.2, II = 3.3, III = 2.4, IV = 1.7, V = 1.1, and VI or F = 0. The multiplication of these by another constant would distribute them proportionately over the interval from 0 − 5, giving I = 5, II = 3.9, III = 2.9, IV = 2.0,

[6] We are indebted to Dr. W. E. Milne for a proof of this proposition.

V = 1.3, VI or F = 0. Thus the traditional weighting of grades for quality already in use at the University of Oregon where an hour of I = 5 points, of II = 4 points, of III = 3 points, of IV = 2 points, of V = 1 point, and of F = 0, is about as close to the statistically determined values as could be expected without resorting to decimals. We have therefore used these traditional weights throughout our study. Marks of "Incomplete" and "Dropped" were ignored. In passing it may be noted that these traditional weights penalize the poor students somewhat and award a small extra premium to the very capable ones.

In all our computations the grade-point ratio—average number of points per hour—has been taken as the measure of student scholastic achievement. Thus it is the reliability of estimates of the *average quality* of the scholastic performance of students which can be inferred from our correlations. We were aware of Wood's statement (12) that the combination of quantity and quality indices results in an index that has greater reliability than either quantity or quality indices taken separately, but we preferred to evaluate the reliability of quality for several reasons. First, quality and quantity of scholastic work are psychologically very different matters which we do not wish to confuse. Second, with 70 per cent of our students self-supporting in whole or in part, inconstant environmental pressures rather than definite personal characteristics are likely to determine the amount of work undertaken much more than the quality of work done. Third, students definitely attempt to fit their scholastic load to the exigencies of economic opportunities which vary from quarter to quarter. Again, it is the quality of work with which we are most concerned rather than the regularity with which credits pile up in the registrar's office. Finally, when grades are weighted as indicated above, partial and multiple correlation shows that average quality (grade points per hour of work taken) is 2.35 times as important a factor in total grade points earned as is the number of hours carried.

Thus our figures for the reliability of college marks were determined by correlating the average quality of two essentially similar samples of each student's work during alternate quarters of the first two years in college. In order to secure reasonable similarity in the abilities and instruction required, the marks of each student were segregated into 19 different divisions of training, and all work done in the odd quarters in each division was compared with all work done in the same division during the even quarters. These divisions were as follows:

1. Architecture.
   Fine arts.
   Normal arts.
2. Botany.
   Zoology.
3. Business administration.
4. Chemistry.
5. Economics.
6. Education.
7. English.
8. Geology.
9. German.
   Greek.
   Latin.
10. History.
11. Journalism.
12. Mathematics.
    Physics.
13. Music.
14. Philosophy.
    Political science.
    Sociology.
15. Psychology.
16. Romance languages.
17. Military science—Men (required).
18. Physical education—Men (required).
19. Personal hygiene—Women (required).
    Physical education—Women (required).

A few departments were consolidated with others of perhaps questionable similarity in order to get enough cases to give reliable determinations of the correlations. The computation for the record of one individual is given below to illustrate the methods used.

Department	Number of hours and grades first year, by terms			Number of hours and grades second year, by terms			Odd terms		Even terms	
	Fall (No. 1)	Winter (No. 2)	Spring (No. 3)	Fall (No. 4)	Winter (No. 5)	Spring (No. 6)	Points over hours	Grade-point ratio	Points over hours	Grade-point ratio
1	2	3	4	5	6	7	8	9	10	11
Education				3 V	3 IV	3 II	6/3	2.0	15/6	2.5
English	4 IV	4 IV	4 III				20/8	2.5	8/4	2.0
History	4 IV	4 IV	4 V	4 III	4 IV	4 III	20/12	1.7	32/12	2.7
Journalism [1]	2 III [1]		2 IV [1]				[1] 10/4	[1] 2.5		
Psychology				3 III 1 inc.	1 III 3 IV	4 III	9/4	2.3	21/7	3.0
Romance languages	4 V	4 V	4 IV	3 V			12/8	1.5	7/7	1.0
Physical education	1 III 1 IV	1 IV 1 F	1 I 1 IV	1 III	1 inc.	1 I	12/4	3.0	10/4	2.5
Weighted mean for all courses								2.1		2.2

[1] Not used because can not be paired.

The results appear by departments and for the university as a whole in the following table.

Reliability coefficients of University of Oregon grades

[Grades expressed in terms of grade-point ratio; based on 2-year records of freshmen entering in 1925]

Department	r_3III* (obtained)	r_6VI* (estimated)	r_1I* (estimated)	r_1I* (obtained)	Number of cases used for obtained r_3III N_3	Average number of quarters paired for r_3III (used for estimated r_1I) n	Number of cases used for obtained r_1I N_1
1	2	3	4	5	6	7	8
Architecture, fine arts, normal art	0.69	0.82	0.52	0.46	119	2.08	51
Botany, zoology	.80	.89	.67	.....	123	1.93	.....
Business administration	.78	.88	.60	.57	101	2.39	60
Chemistry	.71	.83	.57	.....	85	1.86	.....
Economics	.71	.83	.61	.....	126	1.56	.....
Education	.68	.81	.60	.....	138	1.42	.....
English	.77	.87	.60	.36	318	2.23	174
Geology	.58	.73	.46	.....	60	1.59	.....
German, Greek, Latin	.90	.95	.79	.....	76	2.38	.....
History	.75	.86	.63	.....	157	1.73	.....
Journalism	.69	.82	.50	.....	48	2.24	.....
Mathematics, physics	.61	.76	.48	.....	90	1.71	.....
Music	.60	.75	.38	.....	62	2.37	.....
Philosophy, political science, sociology	.56	.71	.46	.....	121	1.44	.....
Psychology	.67	.80	.57	.....	153	1.50	.....
Romance languages	.79	.88	.61	.50	261	2.48	176
Military science	.59	.74	.39	.26	134	2.22	66
Physical education:							
Men	.67	.81	.43	.48	181	2.76	150
Women	.61	.76	.35	.27	212	2.89	195
Total for men	.85	.92	.66	.64	184	3.00	184
Total for women	.90	.95	.76	.65	212	3.00	212
Total for all students	.89	.94	.73	.67	396	3.00	396
PE_r (probable error)							
For departments	.015-.058	.010-.047	.021-.060	.038-.077	.....	.....	.....
For totals	.007-.013	.004-.008	.014-.024	.019-.029	.....	.....	.....

* r_3III = coefficients of correlation between (*a*) average grade-point ratios for courses taken in 3 odd quarters and (*b*) average grade-point ratios for courses taken in 3 even quarters.
r_6VI = estimated correlation which would be obtained if grade-point ratios for 6 odd and 6 even quarters were available.
r_1I (estimated) = estimated correlation which would be obtained between (*a*) grade-point ratio for 1 quarter and (*b*) grade-point ratio for the corresponding quarter in the following year.
r_1I (obtained) = actual coefficient of correlation between (*a*) grade-point ratio for 1 quarter and (*b*) grade-point ratio for the corresponding quarter in the following year.

The direct comparison of the reliability of grades in different departments, from this table, is probably not justified for two reasons. First, the range of grades assigned is not the same for all departments, and correlations increase with increased range.[7] Second, it is always possible that the apparently high reliability of grades in a given department is spurious because instructors make a practice of exchanging opinions about their best and poorest students and hence produce an agreement in marking, throughout the department, based more on the departmental reputation of the student than on his actual performance in different courses.

We have used the Spearman-Brown "prophecy" formula to estimate the correlation of six quarters of college work with six similar quarters, i. e., the reliability of grade-point ratio for lower division

[7] We expect to make corrections for this factor in a later report.

work as a whole. Likewise we have estimated the correlation of grade-point ratio for a single quarter with another similar quarter, i. e., the reliability of grade-point ratio for a single quarter. Wherever the number of cases available warrants it, we have computed the actual correlation of the work done in different departments and in the university as a whole during the winter quarter of 1925–26 and the next winter quarter of 1926–27.

For the university as a whole three things are rather outstanding. (1) Women are more consistent in their scholarship than men. Three quarter grade-point ratios correlate r_3III = 0.85 for 184 men, and r_3III = 0.90 for 212 women. Since this difference is more than twice the standard deviation of such differences, the probabilities are 98 in 100 that it is a real difference, i. e., not due to chance errors in sampling. (2) The reliability of cumulative estimates of average scholarship from various departments is surprisingly high. Low correlations between test scores and college grades are often attributed to the low reliability of the grades. This is reasonable enough where the correlations are with grades in single courses or even with single quarter grades. But low correlations with averaged grades for three or more quarters of college work can not be explained away by disparaging the essential consistency of the grades. The relative accuracy of college marks within a given college group is quite certainly equal to that of our best psychological tests if fairly independent estimates of scholarship are averaged for as much as two years. (3) Published determinations of the reliability of college grades have been computed on such different bases that comparison is difficult. Toops (10) got reports on the correlation of marks for successive semesters in 17 colleges. The average was $r = 0.66$. McPhail, Kornhauser, Cleeton, Crawford, and Wood (1, 2, 5, 6, 13) have reported very similar findings. At Oregon we found a correlation of $r = 0.67$ between grade-point ratios for two similar quarters a year apart. We therefore venture to consider our findings typical of college marks in general. Correlations between the marks of different quarters (and probably of semesters also) understate a little their actual reliability as measures. The difference is surprisingly small, r_1I (estimated) = 0.73 instead of r_1I = 0.67, for a single quarter of lower division work. Hence, in general, correlations between excellence of scholarship in various terms approximate, but rather definitely understate, the degree to which college marks succeed in identifying individual differences in attainment.

BIBLIOGRAPHY

(1) Cleeton, G. U. The predictive value of certain measures of ability in college freshmen. Journal of Educational Research, 15:357–370, 1927.

(2) Crawford, A. B. Forecasting freshman achievement. School and Society. 1930, 31:125–132.

(3) GIFFORD, W. S. Does business want scholars? Harpers, 1928, 156:669–674.
(4) KELLEY, T. L. Statistical methods. New York, The Macmillan Co., 1924 edition, Section 29.
(5) KORNHAUSER, A. W. Test and high-school records as indicators of success in an undergraduate school of business. Journal of Educational Research, 1927, 16:342–356.
(6) MACPHAIL, A. H. The intelligence of college students. Baltimore, Warwick & York, 1924.
(7) RUCH, G. M. The improvement of the written examination. Chicago, Scott, Foresman & Co., 1924. p. 40–64
(8) SPENCE, R. B. The improvement of college marking systems. Teachers College Contributions to Education, No. 252, 1927. 75 p.
(9) STARCH, DANIEL. Educational measurements. New York, The Macmillan Co., 1916. pp. 3–19.
(10) TOOPS, H. A. The status of university intelligence tests in 1923–24 Part II, Journal of Educational Psychology, 17:110–124, 1926.
(11) WOOD, BEN D. Measurement in higher education. Yonkers-on-Hudson, & New York, World Book Co., 1923. 337 p.
(12) ——— p. 133.
(13) ——— p. 98, 131–133.

IMPROVEMENT OF THE ESSAY TYPE EXAMINATION

By R. W. LEIGHTON [1]

In this discussion of the improvement of the essay examination the term essay is used as an inclusive term to designate all examinations which leave the form and the content of the answers to the student taking the examination. This definition of the term is intended to exclude all forms of the so-called new type or objective examination, and it is intended to include all forms of written examinations from those which call for the development of a topic or the writing of a composition to those which call for a bare enumeration of facts. Likewise, these examinations may call for the exercise of any particular mental activity the examiner may wish to invoke.

For some time, but particularly since the advent of the new type examination, the essay examination has been vigorously attacked as an inaccurate measure of student achievement. Most studies of such examinations show that this criticism is usually warranted, yet the essay type is still the one most often used in an institution of higher learning offering such courses as those which are offered here at the university, probably because of its extreme flexibility in adaptation to the measurement of the ability of a student to exercise different mental activities in the various subject fields. Whatever the reason for its popularity may be, the fact that it is the measuring device most often used warrants a careful study of its weaknesses; hence the purpose of this paper is to offer the result of some study of two of these weaknesses.

There are only two weaknesses of the essay examination which materially lower its value as a measuring device. These are: First—its subjectivity, or the inability of different judges rating papers written for such examinations to agree as to the value of these papers as evidence of successful achievement, and second—their low capacity for sampling the student's knowledge in a given subject field. In other words, the number of questions is so limited that students do not have equal opportunity of showing their achievement. Certainly, either or both of these weaknesses make the essay examination a very poor measure when they are present to any great degree.

The new type or objective examinations owe most of their success to the extent to which they eliminate these factors and any improvement of the essay must also begin with their elimination.

[1] R. W. Leighton, secretary of the Committee on Improvement of College Teaching, University of Oregon. Ph. D., University of Oregon, 1932.

During the last three years essay examinations have been studied here at the university whenever it was possible to get two or more judges to rate the papers. These judgments were then correlated in order to determine a mathematical coefficient which could be used to represent the reliability of the judgments or ratings. The first 20 examinations used were drawn from four different departments which used essay tests and the judgments showed correlations ranging from $r = 0.36$ to $r = 0.65$. Four of these r's were below 0.5 and the other 16 ranged from 0.5 to 0.65.

Study of these examinations revealed the fact that a large number of the variations in the scores assigned by the different judges were due to two things, namely, lack of agreement as to what the question asked for, and lack of agreement as to what the answer should be. In the one case this was due to poorly worded questions and in the other to incompetence of the judges. Graduate students had acted as judges in most cases and their judgments did not agree with the judgment of the instructor usually because they were prone to assume that one answer and one only could be correct, even for questions which involved controversial points. There seemed to be no reason to suppose that such variations could not be eliminated nor did there seem to be any reason for classing them as factors of subjectivity.

The next step undertaken was to make similar comparisons in normal situations in which these factors had been minimized as much as possible, and by this means attempt to determine an index of the lowest point to which subjectivity alone need reduce the reliability of judgment, also to determine an index which would represent the greatest possibilities for the reliability of scoring essay examinations. These indices, if they could be established, would give a rather definite picture of the operation of the subjectivity factor. Accordingly a test in philosophy was chosen as the most subjective test to be found among all term tests, and a test in plant biology was chosen as the least subjective essay test it was possible to find.

The philosophy test was one given as a final term test. It required two hours to finish, but no time limit was set, and it contained 10 questions which were to be answered separately. There were 32 papers to be rated. The test called for no factual material for purposes of evaluation; instead the judges looked for such things as revelation of the extent of a student's reading with regard to his ability to choose his readings wisely, and his ability to grasp philosophical problems. These and other equally subjective factors made up the criteria for rating the papers.

No single questions were scored or weighted on any paper, but each paper was scored as a unit, as worth a 1, a 2, a 3, a 4, a 5, or a failure. The judges were the instructor handling the class and an instructor from another department who studied extensively in the field of philosophy.

Little opportunity was left for ambiguity in the questions, and the instructor handling the class very carefully explained to the second judge just what the criterion of judgment should be. Both held as closely to this criterion as they could while rating the papers.

The following conditions, then, existed for this comparison of judgments or scoring. The examination was designed to measure what are commonly called extremely subjective factors.

The criteria by which the papers were to be judged was set up as carefully as possible.

The students were known to one instructor and were not known to the other, so that any so-called "halo" effect which is present in a normal situation was present and operative in this case.

The number of students was small and the range of grading was narrow.

Two competent judges rated the papers, but the instructor handling the class was necessarily the more competent of the two.

The correlation between the resulting grades was $r = 0.63$, which is a rather respectable figure when it is remembered that this situation was picked as the one in which the greatest variation of judgment was to be expected, and it may be assumed to be the lowest point to which subjectivity need lower the reliability of judgment in the case of these examinations.

The second examination used was an examination in plant biology. This examination consisted of five questions designed to measure specific knowledge acquired in the laboratory part of the course. Each question counted as 20 points. The papers were rated first by five laboratory assistants, each of whom rated one question and one only. The examination was prepared by the instructor in charge of the course, and the instructions as to how each question was to be rated were carefully stated by this instructor. Each judge was required to abide by these instructions. This was the usual procedure in that course, so it was used because it was the least deviation from normal procedure and because it was an excellent method of obtaining accuracy of judgment. In order to obtain a coefficient of the reliability of these judgments the papers were given to a graduate student who was not connected with the course and who knew none of the students from the standpoint of their classroom work. This second judge rated all the questions on all the papers but followed the same instructions given to the first judges, and also rated the questions one at a time for all papers; that is, she rated all answers to question 1 before rating any answers to question 2.

The following conditions existed then for this comparison of judgments:

The examination was designed to measure what are commonly considered very objective factors, yet the form of each answer was left to the student.

The criteria for judgment were definitely explained to the judges.

The students were known to the judges in one case and not in the other.

There was a large number of students involved and a wide range of scoring.

The correlation found was $r = 0.90$, which was the highest correlation of any kind found in the study of both essay and objective examinations used as final term measures. It may be assumed to be an index of how high the reliability of judgment may go in the case of the examinations studied.

These two examinations represent the greatest range for the effect of subjectivity upon the reliability of judgment that could be found among the term tests of the essay type. And the total range lies between the correlation coefficients 0.63 and 0.90.

The results described so far offer rather definite evidence of the following: (1) That the range of the effect of subjectivity in scoring essay tests is not so great as is often supposed. (2) That the reliability of judgment increased to a marked degree under the following conditions—when the factors to be measured are carefully determined before writing the examination; when the questions used are carefully planned to meet the requirements of the examination and of clear statement; and when the criteria for judgment are carefully set up.

The extra work imposed upon the instructors so far as actual assigning of grades was concerned was not considered by them to be significant.

Another illustration will show rather well the influence of these factors in improving the reliability of judgment in scoring this type of material. It happened that 115 compositions had been written by entering freshmen. In this case the choice of topics for these compositions was not well controlled and the judges were simply asked to rate them in an order of merit as evidence of the ability of the student to undertake college English. The judges were instructors handling the English courses involved. In this case, therefore, there was some ambiguity in the assignment and no criteria of judgment were set up; instead each instructor rated according to his own ideas. The resulting correlation was low as such correlations may be expected to be when careless procedure is permitted—the r was 0.36. It was desirable to use such compositions as the criterion in another study so one of the instructors worked out a scale for their measurement. One hundred and forty-two people were then asked to write on a given topic which was carefully chosen and carefully assigned, in order that each student might have nearly equal opportunity. The compositions were then carefully rated on the basis of the new scale by two high-school teachers of English, one of 10 years experience and one of 5 years experience. The correlation of the resulting judgments was 0.88.

In this case we have again a great increase in agreement resulting when the directions for writing the compositions were accurately planned and when a definite criterion for judgment was used by the judges. There was no reason to believe that there were any great differences in the abilities of the four judges involved.

It will probably be pointed out that Starch and Elliott, Ruch and others found much different results than those just quoted and it would be well to keep in mind certain things in connection with those studies. They experimented with one paper and a group of judges while these studies used many papers and a few judges, also their conclusions are based upon the numerical ratings on the one paper used. Such ratings may have some value but the value of the ratings would have been much more significant if several papers had been graded and then the ratings compared to see how well the teachers had agreed as to which papers were best, next best, etc., for normally grades are determined by comparative value rather than by percentage values. Probably a far different story would have been told had this been done.

Second, there is no evidence that a very definite criterion for judgment was offered. These same experimenters were later very careful of that factor when building objective tests. Their keys to objective tests are carefully worked out and rigidly adhered to in scoring all such tests.

The findings offered here appear when the judgments are directed and controlled in the same way when the essay test is involved that they are directed and controlled when the new type test is involved. In fact the procedure was deliberately copied from the procedure used with the new type tests. In every case the instructor in charge of the work decided what was to be measured, wrote questions which in his judgment would measure it, and, finally, definitely outlined a key or criterion by which the results were to be judged. The instructor's judgment is always final in this way for ordinary term tests, no matter which type test is used.

The second weakness of the essay test which was mentioned earlier (the matter of inadequate sampling of a student's knowledge in a given field), does not apply to all essay tests. For example, it does not apply in the case of the second group of compositions described, neither does it apply to any extent to the philosophy examination, for the students were given opportunity in both these cases to express themselves concerning subjects with which they had to be familiar if they were to be considered prepared at all.

When the sampling factor does apply (as it did in the biology test quoted) it is difficult to obtain a coefficient which represents its actual effect upon results. We may assume that the correlation of chance halves of the test (as represented by odd *v.* even numbered questions) can not be high if great errors of sampling are present. But this

correlation is complicated by the reliability of the judgment in rating the test and also by nearly all other weaknesses that may be present. However, if a situation can be found in which these other weaknesses are reduced to a minimum, the correlation of chance halves of the test should be affected chiefly by sampling. On this assumption chance halves of the biology test mentioned earlier were correlated and the resulting r, when stepped up by use of the Spearman-Brown formula, was 0.87 so it would seem safe to assume that adequate sampling is possible when care is exercised in handling the examination. A large number of other essay tests were treated in the same manner and gave r's ranging from 0.09 upward, with the median somewhere in the 0.50's. However, seven carefully handled examinations were available which gave r's ranging from 0.67 to the one quoted before, namely, 0.87. As pointed out, the error of sampling is exaggerated when this method of measuring that factor is used, yet it is about as good evidence as can be obtained. Still further study has shown that these correlations of chance halves of essay tests increase materially when a large number of questions, requiring only short answers, is used in these tests, a fact which need only be mentioned to be self-evident.

This study has not been an attempt to treat comprehensively all of the factors which limit the value of the essay examination, but rather an attempt to deal with two of its principal weaknesses. Its other weaknesses are weaknesses of most tests, whatever the type, and should be handled as weaknesses of examinations in general.

SUMMARY

The studies described give evidence of the following:

1. Much of the inaccuracy of rating essay examinations heretofore credited to subjectivity has been due to carelessness in writing questions and in setting up criteria by which answers to these questions were to be judged.

2. The effect of subjectivity upon scoring is not so great as is commonly supposed and it need not reduce the reliability of judgment below a reasonable limit. In some cases it becomes of negligible importance.

3. The essay examination yields readily to the techniques of improvement which are used to improve new type examinations, and experience with the studies described in this paper indicates that greater results are obtained for a small amount of work in the case of the essay examination than for the same amount of work in the case of the new type.

4. The sampling error is not present in many types of essay examinations and when it is present its effect is reduced to reasonable limits by care in selecting questions or by using a large number of short-answer questions.

GROUP II.—STUDENT PERSONNEL STUDIES

AN EVALUATION OF CERTAIN TESTS AND INFORMATION FOR PREDICTING SUCCESS IN NORMAL SCHOOL

C. C. UPSHALL [1] and HARRY V. MASTERS [2]

The problem of this paper is to evaluate certain tests and information for the purpose of predicting the following: (1) Average first quarter grades; (2) average grades during the whole undergraduate period; (3) practice teaching success; (4) whether or not the student will graduate; (5) whether or not the student will receive a position through the Appointment Bureau of the State Normal School at Bellingham; (6) success in teaching in the field during the first semester after graduation.

The Bellingham State Normal School gives eight tests to all entering students. These tests with working time limits are:

	Minutes
1. Thorndike examination for high-school graduates	60
2. History	25
3. English usage	20
4. Arithmetic reasoning	12
5. Arithmetic computation	12
6. Geography	25
7. Spelling	50 words
8. Penmanship	Two 4-minute tests

The Thorndike examination for high-school graduates is used in the grading system of the school. The tests in English usage, arithmetic reasoning, arithmetic computation, and spelling are used to measure the student's ability in these fields. A minimum score (this score is a point minus one-half sigma below the mean of the entering group) must be attained before a student is permitted to do practice teaching. At the present time only three retests are allowed in any one subject. If, after the third retest, the student has not attained the minimum requirement he is advised that he can not graduate from the Bellingham State Normal School. The history test is not used directly at present. It is still in an experimental stage. The

[1] C. C. Upshall, director of the Bureau of Research, State Normal School, Bellingham, Wash. B.A., University of British Columbia, 1923; Ph.D., Columbia University, 1929. He was formerly instructor in the International Institute of Teachers College, and statistician for the New York Commission on Ventilation.

[2] Harry V. Masters, associate director of the Bureau of Research, State Normal School, Bellingham, Wash. B.A., Western Union College, 1924; M.A., University of Iowa, 1925, Ph.D., 1927. Publication: "A Study of Spelling Errors," *University of Iowa Studies in Education*.

geography test has been dropped because it seemed to duplicate certain of the other tests. If a student does not attain the minimum score in penmanship he must take a course in this subject before he is allowed to do his practice teaching.

The reliability of these tests is quite high. Table 1 gives the reliability coefficients for each of the tests with the exception of spelling. The reliability of all the tests with the exception of the Thorndike examination was computed by means of the split-halves technique and corrected by means of the Spearman-Brown prophecy formula. The reliability given for the Thorndike examination is taken from Wood (1).[3]

TABLE 1.—*Reliability coefficients of the tests given to students entering the State Normal School at Bellingham*

Test	r	P. E.	n
1	2	3	4
Thorndike examination	0.85		
History	.95	0.01	197
English usage	.92	.00	436
Arithmetic reasoning	.72	.03	150
Arithmetic computation	.82	.02	150
Geography	.89	.01	405

The majority of the grades which are given in the Bellingham State Normal School are based almost entirely on objective tests which are made by the teachers and scored under the direction of the Bureau of Research. The reliability of the composite scores based upon these objective tests has been computed for several courses. The reliability of these composite scores ranges from 0.24 to 0.96. The median reliability is 0.85. Each instructor, of course, has the liberty of making changes, based on additional information, in the distribution derived from the objective tests. In some subjects very few objective tests are used. It follows then that the reliability of the grade in the typical course will probably be something less than 0.85. On the other hand, average first quarter grades will have a reliability that is superior to the grades given in a single course.

The first problem in this study was the selection of a group from which a regression equation could be derived to predict average first quarter grades. A hundred students, selected at random from the freshmen who entered in the fall of 1927, were chosen. All the tests, with the exception of penmanship, were used in an endeavor to predict the average first quarter grades of this group of 100. These tests were the Thorndike examination, history, English usage, spelling, arithmetic reasoning, arithmetic computation, and geography. In addition the age of the students at high-school graduation

[3] Numbers in parentheses refer to "Bibliography," p. 32.

was used. Table 2 gives all the intercorrelations for these tests with average first quarter grades.

TABLE 2.—*Intercorrelations for the 1927 group*

Variable	1	2	3	4	5	6	7	8	9
Average first quarter grade		−0.43	0.68	0.57	0.36	0.44	0.54	0.51	0.54
Age at high-school graduation			−.30	−.34	−.19	−.44	−.20	−.24	−.26
Thorndike score				.46	.51	.35	.59	.55	.56
History					.33	.31	.34	.29	.77
English usage						.34	.21	.30	.29
Spelling							.19	.32	.36
Arithmetic reasoning								.63	.51
Arithmetic computation									.35
Geography									

It was found that of these factors only the Thorndike examinations, history test, and age of high-school graduation were significant in predicting average first quarter grades. The others added practically nothing to the prediction. The regression equation, using these three predictive factors, is—average first quarter grade = −0.014 times age of high-school graduation in months + 0.036 times the score on the Thorndike examination + 0.012 times the score on the history test + a constant of 0.26. Table 3 shows the means, sigmas, and the regression equation weights for the four variables used in the above equation.

TABLE 3.—*Means, sigmas, and regression equation weights of the variables used in the prediction of average first quarter grades (1927 group)*

Variable	Mean	Sigma	Weight
1	2	3	4
Average first quarter grades	2.2	0.8	
Thorndike examination	116.6	10.8	0.036
Age of graduation from high school	18–3.8	10.8	.014
History	73.7	18.3	.012

The coefficient of multiple correlation from which the regression equation is derived is 0.762. The standard error of estimate is 0.5 of a grade point.

In order to find the value of this equation in predicting average first quarter grades for a new group, all the students who entered as freshmen in the fall of 1928 were chosen for investigation. The average first quarter grade of each student was predicted and the correlation between actual average first quarter grades and predicted first quarter grades was computed. This coefficient of correlation is 0.71 ± 0.02. It is seen that there is a decrease of 0.05 when the regression equation based on the scores of the 1927 group is applied to the 1928 group. This coefficient of correlation may be interpreted as being 30 per cent better than chance.

There were 275 students who entered as freshmen during the fall quarter of 1928. Of these 125 were graduated at the end of the spring quarter 1930. The coefficient of correlation between predicted average first quarter grades and actual average first quarter grades of these 125 students is 0.61 ± 0.04. For this selected group there is a decrease of 0.10 point in the correlation from that found when the 275 students were used. However, as is shown in Table 4, the standard deviation of the 125 group is only 0.47 whereas the standard deviation of the 275 group is 0.74.

TABLE 4.—*Coefficients of correlation between predicted grades and actual grades*

Predicted score	Average first quarter grades	Sigma of predicted scores
1	2	3
1927 group	0.76±0.03	----------
1928 all entrants	.71±0.02	0.74
1930 graduates	.61± .04	.47
1930 graduates	[1] .63± .04	.47

[1] Average grade during six quarters.

The predicted average first quarter grades for the group of 125 were correlated with the actual average grades received during the entire period of attendance. This coefficient of correlation is 0.63 ± 0.04. The standard deviation is 0.47.

In order to determine the relationship between average first quarter grades and the ability of the student to graduate the biserial r technique was used. Table 5 gives this relationship. Table 5 also gives the biserial r between graduating and predicted first quarter grades and between graduating and the Thorndike score.

TABLE 5.—*Biserial coefficients of correlation between graduating and average first quarter grades, predicted first quarter grades, and Thorndike scores*

	Graduating or not
Average first quarter grades	0.51±0.04
Predicted first quarter grades	.37± .05
Thorndike score	.30± .05

Average first quarter grades are distinctly better than either the Thorndike score by itself or the predicted first quarter grades based on the Thorndike score, history score, and age of graduation from high school in estimating whether or not a student will graduate from the Bellingham State Normal School. However, none of the coefficients is sufficiently high for use in predicting an individual case. The predicted first quarter grade and the Thorndike score are practically useless even for the prediction of group achievement.

The tests have proved fairly satisfactory for predicting average first quarter grades. The average grades for the entire course of

those who graduate are reasonably well predicted. The next question to be discussed is the value of the tests for predicting the rating students will receive in their practice teaching. Table 6 shows the coefficients of correlation between practice teaching grades and each of the tests that were used for predicting average first quarter grades. In addition the coefficients of correlation between practice teaching and (1) teaching success in the field during the first semester after graduation, (2) average first quarter grades, (3) estimated first quarter grades, and (4) all the tests combined are shown.

TABLE 6.—*Zero order coefficients of correlation existing between practice teaching and certain other variables and between teaching success and certain other variables*

	Practice teaching	Teaching success
1	2	3
Practice teaching		0.27±0.06
Teaching success		
Thorndike	0.27±0.06	
Average first quarter grades	.33± .06	−.06± .06
Predicted first quarter grades	.45± .05	.02± .06
All tests combined	.36± .06	.03± .06
Age of high-school graduation	.44± .05	.05± .06
Arithmetic reasoning	−.23± .06	−.16± .06
Arithmetic computation	.30± .06	.14± .06
English usage	.23± .06	.06± .06
History	.13± .06	−.05± .06
Spelling	.23± .06	−.04± .06
	−.23± .06	.06± .06

The highest coefficient of correlation is between practice teaching and average first quarter grades. This coefficient is 0.45 ± 0.05. The coefficient of correlation between practice teaching and predicted first quarter grades is 0.36 ± 0.06. When all the tests are combined into one score and this score is correlated with practice teaching, the correlation is 0.44 ± 0.05. It seems that none of the tests adds anything to the prediction of practice teaching after average first quarter grade has been used. The coefficient of alienation of 0.45 is 0.89. This means that practice teaching may be predicted from average first quarter grades with an accuracy only 11 per cent better than chance. Our tests do not predict practice teaching grades nearly as well as they predict average first quarter grades.

Perhaps this lower correlation may be due in part to a low reliability in the practice teaching ratings. While the reliability of practice ratings can not be determined with complete accuracy, it is possible to estimate, within certain limits, what this reliability is. The students who do their practice teaching are given a letter grade rating by their home-room teachers. They are also given a letter grade rating by a supervisor. These ratings are supposed to be made entirely independently. The teacher is obliged to rate each student on a rating scale which has been evolved by this institution. The

supervisors also rate the student independently on this same rating scale. The coefficients of correlation between the ratings given by the teachers on this rating scale and ratings given by the supervisor will give an indication of the reliability of the practice teaching grade. Another measure of reliability of the practice teaching grade is the correlation between the practice teaching grade given by the teacher and that given by the supervisor. Two other indications of reliability may be computed, i. e., the coefficient of correlation between the teacher's rating and the teacher's grade and the coefficient of correlation between the supervisor's rating and the supervisor's grade. Table 7 summarizes these coefficients of correlation.

TABLE 7.—*Coefficients of correlation indicating the reliability of practice teaching grades and ratings*

Factors	1	2	3	4
Supervisor's rating		0.92±0.01	0.80±0.03	
Supervisor's grade	0.92±0.01			0.88±0.02
Teacher's rating	.80± .03	.88± .02		.90± .02
Teacher's grade			.90± .02	

The lowest coefficient of correlation is that between the teacher's rating and the supervisor's rating. This is 0.80 ± 0.03. The highest coefficient of correlation is between the supervisor's rating and the supervisor's grade. It is 0.92 ± 0.01. Since the final teaching grade is the average of the teacher's grade and the supervisor's grade, it is to be expected that the coefficient of reliability will be slightly higher than those reported here. The reliability of the practice teaching grade is as high as the reliability of the objective tests. Naturally a little, if not a large amount of this agreement, is due to similar standards of judgment and similar educational philosophies. This agreement would not be expected between superintendents and principals in the field.

The maximum correlation that could be expected between the practice teaching grade and the tests in the light of the reliability of the tests and the reliability of the teacher ratings of practice teaching would be in the neighborhood of 0.90. The correlation of 0.45 between average first quarter grades and practice teaching is far from being as high as the reliability of the tests would allow.

The rating scale which was used by the teachers and supervisors for indicating the quality of practice teaching was sent to the principals and superintendents who employed the graduates. Ratings by the superintendents and principals were returned for approximately 80 per cent of the students who received positions through the appointment bureau since June, 1930. The coefficients of correlation in Table 6 between teaching success and the other variables are based

on 108 cases for whom these field ratings were secured. It is seen from Table 6 that the most valuable single indication of success in the field is the rating given for practice teaching. This coefficient of correlation, however, is only 0.27 ± 0.06. It is barely significant, being only four and one-half times its probable error. In terms of accuracy of prediction this is only 4 per cent better than chance. None of the other factors that were used to predict teaching success does so reliably. All the coefficients of correlation are within three times the probable error of the coefficients. It would appear then that, so far as predicting teaching success is concerned, the tests and information used in this study are worthless with the possible exception of the practice teaching rating.

A measure of the reliability of the ratings given by the superintendents and principals should be secured in order to estimate the amount of correlation that might be expected between teaching success and the various other factors. An indication of this reliability is being secured at the present time by means of another rating which will be returned during May of this year. Until this measure of reliability is secured it is impossible to tell how high a correlation might be expected between practice teaching and success in the field.

It is interesting to compare the results achieved in the Bellingham Normal School with the results achieved by other institutions. Ullman in the Journal of Educational Administration and Supervision for November, 1930, reports the results which he got in an investigation of 116 graduates who were teaching in junior or senior high schools. Table 8 gives the zero order coefficients of correlation reported by Ullman for certain factors which are comparable with those which were obtained in this study.

TABLE 8.—*Certain zero order coefficients of correlation reported by Ullman*

Factors	1	2	3	4	5	6
Practice teaching		0.36	0.24	0.26	0.46	0.22
Teaching success	0.36		.15	.30	.30	.20
Brown psychological examination	.24	.15		.30	.39	.29
Academic marks	.26	.30	.30		.55	.81
Professional marks	.46	.30	.39	.55		.60
Major subject marks	.22	.20	.29	.81	.60	

He reports a coefficient of correlation of 0.36 between practice teaching and teaching success, a coefficient of 0.15 between the Brown Psychological Examination and success in the field, and a coefficient of correlation of 0.24 between the Brown psychological examination and practice teaching. Practice teaching correlates to the extent of 0.46 with professional marks whereas it correlates only 0.26 and 0.22 with academic marks and major subject marks, respectively. Ullman reports a coefficient of correlation of 0.30 between teaching success in

the field and professional marks. This coefficient of correlation is considerably higher than the one obtained in this study, i. e., 0.02. The difference may be accounted for, in part, by the fact that in rating junior and senior high school teachers more emphasis is placed on comprehension of subject matter than in rating elementary school teachers. In general, however, it may be said that the results reported in this paper agree with the results reported by Ullman. At present no single test or rating seems to predict teaching success at all reliably.

TABLE 9.—*Biserial coefficients of correlation between receiving a position after graduation and certain other variables*

	Position
First quarter grades	0.48 ± 0.05
Predicted first quarter grades	.23 ± .05
Supervisor's rating	.35 ± .05
Thorndike	.26 ± .05
Mean of first quarter grades and predicted first quarter grades	.45 ± .05

Table 9 shows the biserial r between receiving a position through the appointment bureau and (1) first quarter grades, (2) predicted first quarter grades, (3) supervisor's rating, (4) Thorndike examination, and (5) the mean of first quarter grades and predicted first quarter grades. The highest coefficient of correlation is again with first quarter grades, i. e., 0.48. Predicted first quarter grades indicate very poorly indeed whether or not a student will receive a position. The Thorndike examination is almost as unreliable. The mean of average first quarter grades and predicted first quarter grades does not increase the coefficient of correlation that is found when first quarter grades alone are used. The supervisor's rating correlates 0.35 with receiving a position through the appointment bureau.

It will be seen that none of the coefficients of correlations reported in this study is of sufficient predictive value to estimate reliably the score that would be obtained by an individual student. In guiding individual students who attend the normal school, it is only possible to indicate their chances of success. It is seldom possible to be certain that a student will succeed or fail although the average first quarter grade of most students can be predicted fairly reliably. Whether or not a student will graduate from the Bellingham State Normal School can be estimated less reliably but still sufficiently well to be useful. Whether or not a student will receive a position through the appointment bureau after graduating can be predicted with some degree of accuracy. With the present reliability and validity of our tests and ratings it is impossible to give the entering student any indication at all of what his rating as a teacher in the field will be.

In order to be able to give information to the students which will be more meaningful to them than the coefficients of correlation reported in this paper, the information summarized in Table 10 has been prepared.

TABLE 10.—*Per cent of those entering the Bellingham State Normal School that graduate, the per cent of those entering that receive positions, and the per cent of those who graduate that receive positions*

Item	Number entering in 1928	Range in grades	Per cent graduating in 1930	Per cent receiving positions in 1930	Per cent of those graduating that received positions
1	2	3	4	5	6
Actual first quarter grade [1]	1	4.0	0	0	0
	52	3.0 to 3.9 inclusive	65	54	82
	111	2.0 to 2.9 inclusive	58	31	53
	102	1.0 to 1.9 inclusive	25	14	54
	9	0.0 to 0.9 inclusive	0	0	0
Predicted first quarter grade	31	3.0 to 3.9 inclusive	65	45	70
	151	2.0 to 2.9 inclusive	55	33	60
	88	1.0 to 1.9 inclusive	25	14	50
	5	0.0 to 0.9 inclusive	0	0	0
Mean of actual first quarter grade and predicted first quarter grade.	35	3.0 to 3.9 inclusive	60	49	81
	141	2.0 to 2.9 inclusive	61	35	57
	93	1.0 to 1.9 inclusive	19	11	56
	6	0.0 to 0.9 inclusive	0	0	0
Thorndike examination	31	A ratings	48	32	67
	68	B ratings	65	44	68
	93	C ratings	47	25	52
	67	D ratings	28	16	58
	16	E ratings	23	15	57
Supervisor's rating	3	4.0			67
	32	3.5 to 3.9 inclusive			78
	75	3.0 to 3.4 inclusive			76
	77	2.5 to 2.9 inclusive			53
	29	2.0 to 2.4 inclusive			48
	2	1.5 to 1.9 inclusive			0

[1] In computing these average grades an A was given a weight of 4; B, weight of 3; C, weight of 2; D, weight of 1, and F, weight of 0. The average grade is the sum of the grades for the courses multiplied by the credit in hours of the course divided by the total number of hours carried.

Table 10 shows the distributions of (1) average first quarter grades, (2) predicted first quarter grades, (3) the mean of average first quarter grades and predicted first quarter grades, and (4) Thorndike scores for the 275 students who entered as freshmen in the fall of 1928. The distribution of the supervisor's ratings of 218 students who graduated during the school year 1929–30 is also given. It gives for each level of achievement, for each of the variables, the per cent of those entering in 1928 who graduated in 1930. In column 5 it gives the per cent of those entering in 1928 who received positions through the appointment bureau, and in column 6 it gives the per cent of those who actually graduated who received positions. As would be expected from the coefficient of correlation between actual first quarter grades and ability to graduate those who receive high actual first quarter grades are more likely to graduate than those who receive low average first quarter grades. Sixty-five per cent of those students who receive average first quarter grades between 3 and 4 were graduated, 54 per cent received positions, and 82 per cent of those that graduated received positions. At the other end of the distribution it is seen that none of those who received average first quarter grades between 0 and 1 was able to graduate. Only 25 per cent of those that received average first quarter grades between 1

and 2 were able to graduate and only 14 per cent received positions. Those students, then, that received average actual first quarter grades below 2, have a relatively small chance of graduating, actually less chance than one chance in four. They have only one chance in seven of receiving a position. However, if these students graduate they have as good an opportunity to receive a position as those that receive average first quarter grades between 2 and 3.

Sixty-five per cent of those who received predicted first quarter grades between 3 and 4 were graduated, but only 45 per cent of this group received positions. Seventy per cent of those that were graduated received positions. None of those students who received predicted first quarter grades between 0 and 1 was graduated. Only one in four of the students who received predicted first quarter grades between 1 and 2 were graduated and only one in seven received positions. Forty-eight per cent of those who received A ratings on the Thorndike examination were graduated and 32 per cent received positions. (A rating of A is equivalent to a score which is between one and one-half and two and one-half sigma above the mean of the entering group. A rating of E is equivalent to a score which is between one and one-half and two and one-half sigma below the mean.) Sixty-five per cent of those who received B ratings were graduated and 44 per cent received positions. Twenty-three per cent of those that received an E rating were graduated and 15 per cent received positions.

Of those students who received a rating between 1.5 and 1.9 from the supervisors none received positions. Forty-eight per cent of those who received ratings from the supervisor between 2 and 2.5 received positions. At the other end of the distribution, of those who received ratings between 3.5 and 4, 78 per cent received positions.

It is possible, as a result of this analysis, to predict with practical certainty by the end of the first quarter that certain students will not be able to graduate and consequently will not be able to receive a position as a teacher. It is possible to tell what the chances of the other students will be of graduating and what their chances of receiving a position will be. There are only about three chances in five of the best students (as judged by actual first quarter grades, or predicted first quarter grades) graduating from the Bellingham State Normal School in the usual period of six quarters. It is not known how many of the students will graduate at a later date. It is possible that some of the better students attend the Bellingham Normal School for only a short period and then transfer to another institution of higher learning from which they were or will be graduated. It is doubtful whether those students who receive low ratings will graduate from this institution or from any other institution of higher learning.

SUMMARY

The facts and conclusions of this paper may be summarized as follows:

1. A coefficient of multiple correlation of 0.76 ±0.03 was found between average first-quarter grades and the combined scores of the Thorndike examination, history test, and the age of graduation from high school.

2. The first-quarter grades of all students who entered in the fall of 1928 were predicted by means of the regression equation obtained from the 1927 group. The coefficient of correlation between these predicted first-quarter grades and the actual first-quarter grades was 0.71 ±0.02.

3. One hundred and twenty-five of the students who entered in the fall of 1928 were graduated in the spring of 1930. The coefficient of correlation between the predicted grades of this selected group and the actual average first-quarter grades was 0.61 ±0.04.

4. The coefficient of correlation between the rating which the student received in his practice teaching and the actual average first-quarter grades is 0.45 ±0.05. This coefficient was not significantly increased by adding to first-quarter grades the combined effect of the entrance tests.

5. The coefficient of correlation between the ratings in practice teaching and the predicted first-quarter grades was only 0.36 ±0.06.

6. The reliability of the practice teaching rating is between 0.80 and 0.90.

7. Average first-quarter grades, predicted first-quarter grades and Thorndike examination have considerable value in predicting whether or not a student is likely to continue in school until graduation and whether or not he will obtain a position through the appointment bureau of the institution. (*a*) Only one in six of those who received D or E ratings on the Thorndike examination was placed by the appointment bureau. (*b*) Sixty-five per cent of the entering group who received actual first-quarter grades or predicted first-quarter grades between 3 and 4 were graduated. (*c*) Fifty-four per cent of the entering group that received actual first-quarter grades between 3 and 4 received positions. (*d*) Eighty-two per cent of the graduates who received actual first-quarter grades between 3 and 4 received positions while only 53 per cent of the graduates who received average first-quarter grades between 1 and 2 or 2 and 3 received positions. (*e*) None of the students who received actual first-quarter grades or predicted first-quarter grades between 0 and 1 was graduated. (*f*) Only one in four of those that received first-quarter grades or predicted first-quarter grades between 1 and 2 was graduated.

8. Ratings of the 1930 graduates who received positions were received from the superintendents and principals. The reliability of these ratings has not yet been determined.

9. The rating in practice teaching is the best single index of the rating which the student will receive in the field. The correlation is very low, i. e., 0.27.

10. All the tests, grades, and information used in this study give very low coefficients of correlation when used to predict the rating which the student will receive in the field.

BIBLIOGRAPHY

(1) Wood, Ben D. Measurement in Higher Education. Yonkers-on-Hudson, New York: World Book Co., 1923. p. 45.

THE SIGNIFICANCE OF PERSONNEL MEASURES AT THE UNIVERSITY OF OREGON

HOWARD R. TAYLOR and CLIFFORD L. CONSTANCE[1]

When a competent engineer measures the potential electric energy of a stream, his results appear to be final and absolute in so far as we are perfectly familiar with, or trustful of, the units used and the mathematical accuracy of their transmutation from second-feet of stream flow to kilowatts or horsepower. In reality, of course, the measurements are purely relative. If you turn to the dictionary for a definition of second-feet or watts or horsepower, you will be rewarded with a translation into other terms, and your search will end only when you encounter terms which you understand well enough to accept without question or, more often, when you have traced out a circular series of terms which ends at the word with which you started. Again, to the naive individual, nothing could be more nonsensical than the engineer's statement of numerical equivalence between the boiling white water on a rapid and the predicted magnetic deflection of a needle on a gage in a power house as yet unbuilt.

Now, the attempt to measure potential college scholarship in terms of personnel data is essentially a comparison of the same sort but considerably more difficult, for two reasons. First, the units used in both sets of measurements are not sufficiently familiar or definite to have a widely accepted meaning. Experts in mental testing are only partially agreed as to the meaning of test scores, and college scholarship in terms of marks or grades is of rather doubtful significance even among those most familiar with the symbols in which such estimates are recorded. Second, the equivalence of each set of units in terms of the other has been very imperfectly established as yet. If we could be sure of the psychologist's mental tests as measures of intellect, then we could determine the meaning of college marks in terms of such intelligence. Or if we accept college marks as good relative measures of the extent to which students have profited from instruction, as indeed many studies of the rough but essential validity of college marks strongly suggest, we can then appraise the tests and other personnel measures in terms of such scholarship. But any gain in our knowledge of either characteristic must be purchased by making one or the other of these assumptions, and then by evaluating the relationship which exists between the two sets of measures. In fact, such

[1] See footnotes 2 and 3, p. 5.

procedure is implicit in measurement of any kind. But in the case of the power-plant engineer, previous laboratory experimentation has already provided a knowledge of the essential relationships between his variables, whereas the psychologist attempting to use personnel measures must continuously work out these relationships for himself while at the same time he collects his data.

Obviously the determination of the relation between personnel measures and the excellence of scholastic achievement would have little merit if the test scores and other measures were themselves utilized as a basis for giving the marks. For this reason, in addition to others equally important, the personnel research bureau has steadfastly restricted the use of personnel data at the University of Oregon, especially for the two groups reported here, to conferences sought voluntarily by individual students and to the administrative analysis of the probable causes of scholastic difficulty after marks in various courses were already assigned. In general, ratings were not released to instructors and were used only under conditions which practically guaranteed the independence of estimates of scholarship from a knowledge of the personnel ratings during the period of this experiment.

In September, 1925, we began collecting personnel information for all students entering as freshmen. Of 454 men and 409 women entering at that time, 62 men (13.7 per cent) and 110 women (26.9 per cent) had completed their work for graduation at the end of the summer session in August, 1929—four calendar years after entrance. In September, 1926, of 455 men and 367 women entering as freshmen, 58 men (12.7 per cent) and 101 women (27.5 per cent) had graduated by August, 1930—four years later. Thus, just less than 20 per cent of those entering the University of Oregon as freshmen may be expected to graduate within the 4-year period following their entrance. It is the records of these 331 students who graduated within four years after entrance which constitute the basic data for this study. In addition we found that 46 men (10.1 per cent) and 27 women (6.6 per cent)—(or 8.5 per cent in all)—of the class entering in 1925 graduated sometime during the fifth year after entrance. It is probably safe to say, therefore, that not more than 30 to 35 per cent of those who enter the University of Oregon as freshmen will graduate from it. Apparently this is about what happens at other reputable State universities.[2]

[2] The University of Minnesota reports in its 1928 volume on "Problems of college education," p. 204, that 24 per cent of the class entering in 1920 had either graduated or transferred to a professional school where they had completed their fourth year of work at the end of the normal 4-year period. Since our figures do not include those members of the 1925 entering class who had transferred to law, medicine, or architecture and completed four years without taking a degree, the 20 per cent who did graduate four years after entrance at Oregon are very nearly equivalent to the 24 per cent reported at Minnesota. Again, Toops and Edgerton, of Ohio State, in their 1929 report of the "Academic progress of students" say, p. 136, that of all the students who enter the university only about 34 per cent will probably graduate. This checks very closely with our estimate of 30 to 35 per cent who will graduate, based on the fact that 28.4 per cent had graduated 5-years after entrance.

Now, if one thinks of the attainment of a degree as the *sine qua non* of university training, this elimination of fully two-thirds of those who enter as aspiring freshmen will seem to constitute a major educational tragedy, mitigated to some extent by the rather small percentage who complete their college work at some other institution, presumably better adapted to their needs and also mitigated by the fact that some of these students have gone on into professional schools, such as medicine, law, or architecture, without taking a general university degree.[3]

On the other hand, if a major function of the university is to furnish selective guidance and assist students in the inevitable trial and error procedures of finding their real aptitudes and interests, extensive elimination may merely evidence the intellectual standards of the university and its efficiency in shifting students out of unpromising endeavors. Moreover, it is absurd to think of college experience which does not eventuate in a degree as time, energy, and money thrown away. It may well be that the benefit from college training, per se, is in many cases as great for those who do not graduate as for a good many of those who do, although in general the prestige and personal satisfaction attached to the degree make it well worth striving for.

But whatever the essential nature of the educational processes leading to graduation—whether they be primarily instructional or selective—we have no scientific basis for advising and guiding students, nor even for reorganizing and improving our scholastic procedures until we know the meaning of our personnel measures in terms of the fundamental objectives of college education which ostensibly, at least, are scholarly achievement in various fields of knowledge. What can the results of an hour or two spent in puzzling out complex mental tasks of an unfamiliar and apparently quite impractical sort possibly foretell about scholastic achievement in say 50 or 60 courses under 20 or 30 different instructors in three or more broad fields of knowledge requiring normally four years of fairly consistent study? To the person unfamiliar with the great variability of college students in mental capacity and unaware of the surprising stability and unitary nature of psychological samplings of general ability, the expectation of any connection at all between such variables seems nonsensical. But that is exactly the reason for resorting to experiment.

The psychological examination of the American Council on Education, which we have given each year beginning in 1925 to all entering freshmen, is a general ability or general college aptitude battery of

[3] According to questionnaire data collected by Mr. Hagstrom, the university editor, in a follow-up of 980 students enrolled in 1928-29, who did not return in the fall of 1929 -30, 20 per cent of those who replied were enrolled in other institutions, and 10 per cent plan to enroll elsewhere, while 30 per cent plan to reenter the University of Oregon. Since only 25 per cent of the 980 circularized replied, it is hard to say just how representative these percentages are of Oregon students in general.

the type commonly called an intelligence test. It samples the ability to do complex mental tasks where previous training as a factor in success has been intentionally minimized. In 1925 it consisted of eight tests for which an hour and a half of working time was allowed. In 1926 only seven tests were used, and it has since been reduced to five tests with an hour of working time without reducing its effectiveness in predicting scholarship. (See following table.)

Evidently the five tests which have constituted the backbone of the battery each year from 1925 to the present give an adequate sample of such general abilities as are scholastically important. The reliability of this battery by the split-halves method for the series of 1925 and that of 1926 we found was $r_1 I = 0.95$, and this value has been extensively corroborated elsewhere (8).[4] But we are concerned here with the stability of such a measure over a long period of time as well as with the extent to which the same abilities are sampled by two similar forms of the test. In April, 1927, we retested with the American Council series of 1926, 93 students who had taken the American Council series of 1925 in September, a year and a half earlier. Thus we sampled the relative stability of such scores after five quarters of university instruction. The correlation of total weighted score on the two independent but similar test batteries was $r_1 I = 0.90$. While the number of cases ($N = 93$) is small, in comparison with our other data, I feel quite confident that the coefficient $r_1 I = 0.90$ does not overstate the extent to which American Council test scores represent a fairly unitary and stable aspect of the potential capacity of college students. In order to facilitate explanation we transmute all our personnel data into percentile ranks (P. R.), indicating the percentage of students entering the University of Oregon as freshmen who made scores lower than the particular score under consideration. The percentile ranks of American Council test scores for our Oregon freshmen, particularly in the upper ranges, approximate very closely the national norms for 1925 based on 16,000 students in 55 different colleges (1). While we find empirically that correlations determined from measures recorded in percentile ranks are slightly lower than where the original measures are used—as is to be expected theoretically—the shrinkage in relationship is negligible and is more than offset by convenience. Thus it is general ability measured in such percentile rank units as we find most convenient to use which we wish to compare with scholarship.

[4] Numbers in parentheses refer to "Bibliography," p. 49.

	Series of—					
	1925	1926	1927	1928	1929	1930
1	2	3	4	5	6	7
r total	0.425			0.484	0.489	0.511
Men		0.440	0.466		.442	.454
Women		.345	.569		.556	.579
N total	189			970	808	800
Men		438	470		420	433
Women		364	392		388	367
Variables	Median P. R.[1] v. grade points	Gross P. R. v. grade average	Gross P. R. v. G. P. R.	Gross P. R. v. G. P. R.	Gross P. R. v. G. P. R.	Gross P. R. v. G. P. R.

[1] P. R. means percentile rank as explained later, and G. P. R. means grade-point ratio as explained later.

But the question of the numerical representation of scholastic excellence is an indispensable preliminary. Instructors evaluate the work of students at the University of Oregon under six catagories. If grade III (slightly above average) and grade IV (slightly below average) be combined into an average group, the scale corresponds to the 5-step scales more commonly used elsewhere. If the grade I were assigned to 5 per cent, II to 20 per cent, III and IV to 50 per cent, V to 20 per cent, and VI or F to 5 per cent, the distribution of grades would approximate the normal curve which has empirically been found to represent very well the actual distribution of grades in several universities where the performance of large groups of students over fairly long periods of time has been studied (2, 3, 6, 7, 10). At Oregon the actual percentages of each grade assigned during representative fall, winter, and spring terms combined were: I = 8.8, II = 23.7, III and IV = 53.5, V = 8.9, and VI or F = 5.0.

In view of the host of diverse and fairly independent factors entering into quality of achievement, the assumption that it is normally distributed is reasonable on a theoretical as well as on an empirical basis.

If now a unit normal distribution be divided into areas proportional to the percentages of students assigned each grade, we can compute the distance above and below the mean where the mid-point of each such area intersects the base line. The relative distances of each such point, in units of variability from the mean, may be considered numerically comparable measures of such differences in the quality of work as instructors can recognize and value. In standard deviation units, these distances corresponding to the various percentages of grades actually assigned at Oregon are I = +1.81, II = +.86, III = +.04, IV = −.69, V = −1.27, and F = −2.40. Since negative weights are awkward to use in actual practice, we can preserve the same mathematical relationships by adding 2.40 to each weight, giving I = 4.2, II = 3.26, III = 2.44, IV = 1.71, V = 1.13, and VI or F = 0. This indi-

cates that the traditional, numerical weighting of grades long used by the registrar's office of I = 5 points, II = 4 points, III = 3 points, IV = 2 points, V = 1 point, and VI or F = 0, is not greatly in error. Actually, in terms of the frequency with which they are assigned, it overweights high grades a little, and underweights the low grades somewhat.

A priori it would seem that grade points (number of hours earned times the weighting for quality) would give the best index of student achievement since it includes both quantity and quality of work. But the university sets up a definite number of hours which each graduate must obtain. Hence it would be necessary to use some such measure as average number of grade points per term if both factors are to count. Now in general, the factors contributing to variability in hours carried are so diverse, e. g., self-support, student activities, health, etc., and the variations in load carried so supervised with reference to the needs of the individual, that we have preferred to measure college scholarship in terms of quality alone. Again quantity and quality of achievement appear to be psychologically very different aspects of a student and certainly the negative correlation between quality and quantity of work implied in the combined index is quite the reverse of the facts. Finally, there is a widespread and growing conviction that quality of scholarship is what needs emphasis in American higher education. Hence total grade points (hours multiplied by the traditional weighting for quality) have been divided by the total number of hours to give average grade points earned per hour for which the student registered. Since VI or F (failure) counts zero points but still counts as hours, all failing grades lower the quality index in the proper proportion. This index of the quality of college achievement which we call the grade-point ratio (G. P. R.) will be used as the measure of scholarship.

Our studies of the reliability of such grade-point ratios at Oregon show that the correlation of three alternate quarters of college work measured in these terms with three comparable quarters of similar work is $r_{3f_3F} = 0.89$ ($N = 396$). Thus we can estimate the reliability of our 4-year grade-point ratios as not less than $r_{12f12F} = 0.97$.[5] Evidently the general factors which underlie pooled estimates of the quality of college work are remarkably consistent and stable. Thus the reliability of averaged grades is much higher than one might expect from the well-known inaccuracies of grades in single courses.

Probably the simplest method of representing the relationship between two variables is to plot a straight line showing the average amount of change in one variable which corresponds to an average amount of change in the other. With test data this is usually done mathematically by computing the correlation coefficient between the

[5] The Spearman-Brown prophecy formula is widely used in this connection and there is considerable empirical evidence of its validity for such data.

two variables. Of course, relationships may exist which can not be represented in such simple linear fashion. Hence, an obtained correlation of $r=0$ does not prove that no relationship exists but merely that no straight line representation of whatever relationship exists is feasible. The extent to which the numerical value of r approaches ± 1 does, however, indicate the extent to which a linear estimate of the one variable is possible from a knowledge of the other. Hence the size of the obtained correlation coefficient between two variables in comparison with the size of its probable error, is indisputable evidence of overlapping factors or identical elements in the two sets of measures. The relationships between test scores and 4-year scholarship at Oregon are as follows.

TABLE 1.—*Four-year study showing correlations between test percentile and four-year G. P. R.—all departments included*

Sex	1929		1930		Both years	
	N	r	N	r	N	r
1	2	3	4	5	6	7
Men	62	0.489	58	0.496	120	0.487
Women	110	.615	101	.532	211	.576
Total	172	.561	159	.489	331	.526

Two things in this table are especially worth noting: (1) Contrary to prevailing belief the behavior of women is somewhat more predictable than that of the men. At least their scholarship can be estimated in terms of test score more accurately than can that of the men—$r=0.58$ as compared with $r=0.49$ for men. (2) According to the most conservative interpretation (9) there is not less than 28 per cent commonality between general ability measured by an hour and a half spent in working at complex mental tasks and the excellence of college scholarship in general over a period of four years. Of course our tests do not pretend to measure preparation, studious habits, nor scholarly zeal except as these may be correlated with general ability. Such characteristics with many others probably constitute the other factors (not more than 72 per cent) in scholarship which do not vary concomitantly with test score. To what extent are these factors also measurable?

Even with the advantage of great accuracy in his separate measurements, no engineer would attempt to gage the potential energy of a stream by a single measurement. He needs a more representative sample of the yearly and year to year ups and downs of the river. Likewise, we need a measure of student preparation and earnestness spread over a period of years. The prevailing practice in regard to

entrance requirements assumes that performance in preparatory schools is prognostic of college performance. Yet due primarily, I think, to variations in the standards of grading in different high schools, average preparatory school grades have usually given very slight evidence only of college success. However, in our personnel research at Oregon we have always considered the evaluation and use of information already on file, such as preparatory school records, quite as important as the collection of additional information such as psychological test scores. Accordingly, we began in 1925 to compute for all entering freshmen an index of the quality of preparatory school work. No doubt a good objective test of high-school knowledge and skill, such as the Sones-Harry or Iowa High School content examinations, would be even more useful, but over a period of six years our index has earned equal weight with general ability test scores in a regression equation for the prediction of college scholarship.

Without describing the several variations in our procedure leading to better prediction and easier computation, the essentials of our method are as follows: We empirically equate the various high-school grading systems and then weight heavily each unit of credit rated in the highest scholastic interval. This of course makes the index depend chiefly on such work as was recognized by the high school as outstanding. This index is also transmuted into percentile ranks for convenience in explanation. Thus it means that such and such a percentage of our entering freshmen had preparatory school records below the one under consideration.

We have no satisfactory method of determining the reliability of this measure as yet. However, the principal of each school was requested on the entrance blank to place each student in the first, second, third, or fourth quartile of the graduating class. If we assume an approximately normal distribution of ability in graduating classes with roughly equivalent means, these categories can be transmuted into numerical measures by assigning the appropriate standard deviation values to the percentage of freshmen falling in each quartile. For these 1925 freshmen, the correlation of the principals' quartile rating with our empirical index of the quality of preparatory school work was $r = 0.62$ for 248 women, and $r = 0.68$ for 174 men, or $r = 0.65$ for 422 cases. For the 1927 freshmen similar computations gave $r = 0.77$ for 320 women, $r = 0.71$ for 372 men, and $r = 0.73$ for 702 cases in all. None of these coefficients can be considered a satisfactory reliability coefficient, but such a coefficient would hardly be higher than $r_{1I} = 0.7$, which would therefore be a conservative estimate.

Fortunately for our purpose preparatory school records are not closely related to general ability test scores. For the 331 graduates of this study the correlation is only $r = 0.44$ ($r = 0.48$ for 120 men, and $r = 0.41$ for 211 women). Thus using the conservative inter-

pretation previously referred to, there may be as little as 19 per cent commonality between these measures of preparatory school work and the general ability test scores. Among many possible explanations, our experience suggests that this is because effort, i. e., docility, dependability, cooperative attitudes, and persistence, are more important factors in preparatory school success than is all-around intellectual ability. Especially does our experience with individual personnel records support such an interpretation. When a student enters with a low test percentile rank (P. R.), but high high-school percentile rank, we usually find a hard-working, serious-minded individual whose difficulties are chiefly slow learning, or inability to appreciate abstract generalizations or to display critical insight. On the other hand, freshmen with high test percentile ranks and low high-school percentile ranks are bad college risks because they have formed habits of loafing and just getting by in place of efficient study habits. Their difficulties are chiefly inadequate preparation and the surprising stability of habits of superficial thinking and of dawdling over scholastic tasks.

The relationships of high-school record to 4-year college scholarship are as follows:

TABLE 2.—*Four-year study showing correlations between high-school percentiles and 4-year grade-point ratios, all departments included*

Sex	1929		1930		Both years	
	N	*r*	*N*	*r*	*N*	*r*
1	2	3	4	5	6	7
Men	62	0.511	58	0.529	120	0.524
Women	110	.528	101	.550	211	.538
Total	172	.514	159	.531	331	.523

In Table 3 the results of combining test percentile ranks with high-school percentile ranks in the prediction of 4-year scholarship are shown:

TABLE 3.—*Four-year study showing correlations between average of high-school and test percentage ranks and 4-year grade-point ratio*

Sex	1929		1930		Both years	
	N	*r*	*N*	*r*	*N*	*r*
1	2	3	4	5	6	7
Men	62	0.587	58	0.591	120	0.588
Women	110	.698	101	.630	211	.664
Total	172	.646	159	.592	331	.620

From Table 2 we may conclude that in spite of the inaccuracies involved in equating high-school marks, an index can be derived from them which is approximately equal in predictive significance over the whole 4-year college period to that of American Council General Ability Test scores. Turning this finding about we may say certainly that the logic of basing admission to college on preparatory school records applies with equal force to psychological test scores.

In two respects at least the latter are superior. (*a*) They are highly objective and hence impartial. Since special training is minimized as a factor in making high test scores, all students are more nearly on the same basis than is the case with preparatory school records with their inevitable variation in standards of grading. (*b*) The psychological test furnishes an adequate sample of general ability in an hour and a half or less, while it takes three to four years of observation in preparatory school at least as at present recorded, to give equal predictions of college scholarship.

From Table 3 the supplementary value of preparatory school records is made clear. When high-school percentile rank is averaged with test percentile rank, the correlation with 4-year college scholarship rises to $r = 0.62$ within the rather restricted range of ability which achieves graduation in that time. Thus in this combined measure we have at least 38 per cent commonality between personnel ratings and college scholarship—a gain of at least 10 per cent in unique factors over either personnel measure alone.

We have also made an extensive analysis of the scholastic significance by departments and schools for each of the five tests in the American Council battery. Our data hardly justify the use of Spearman's mathematical techniques for determining to what extent the relationships found between separate tests, preparatory-school records, and 4-year college scholarship may be thought of as due to the presence of a single general factor plus many specific factors. But there is a strong suggestion that the burden of prediction for each of our measures is carried by some unitary set of factors which is sampled again and again rather than by the teaming up of independent groups of factors constituting unique traits. All our departmental correlations are positive and appear to be in hierarchical order ranging down toward zero for certain tests in certain departments, but never being significantly negative with scholarship in any department. Again the intercorrelations of each separate measure with any of the others and with the criterion of 4-year scholarship (grade-point ratio) is always positive as appears below.

TABLE 4.—*Four-year study showing intercorrelations of 8 dependent variables with the criterion 4-year scholarship (1925–1929 and 1926–1930 groups combined)*

Men.....N=120; 62 on analogies
Women.....N=211; 110 on analogies
Total..N=331; 172 on analogies

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1=4-year grade-point ratio		0.524	0.487	0.461	0.397	0.304	0.298	0.338	0.719
		.538	.576	.452	.415	.446	.371	.489	.811
		.523	.526	.435	.426	.334	.265	.442	.780
X_2=high-school percentile rank			.480	.409	.303	.215	.386	.254	.433
			.409	.421	.345	.074	.348	.311	.482
			.436	.417	.291	.127	.347	.284	.461
X_3=test percentile rank				.737	.710	.721	.585	.681	.483
				.696	.697	.706	.705	.724	.528
				.713	.613	.710	.646	.691	.504
X_4=completion percentile rank					.506	.381	.442	.541	.438
					.480	.371	.400	.539	.398
					.415	.373	.418	.523	.401
X_5=artificial-language percentile rank						.491	.338	.494	.414
						.310	.396	.441	.400
						.336	.177	.466	.397
X_6=analogies percentile rank							.310	.425	.255
							.388	.401	.478
							.345	.407	.398
X_7=arithmetic percentile rank								.157	.314
								.407	.340
								.250	.282
X_8=opposites percentile rank									.395
									.426
									.418
X_9=first-term grade-point ratio									

But when these measures are teamed up by multiple correlation for the prediction of 4-year scholarship, the unique contribution of each separate measure diminishes considerably as the overlapping general factors included in each measure are appropriated by those most heavily saturated with them at the start. This is well shown by the partial correlation coefficients of the fifth order for each measure with the criterion.[6] $r_{12.45678}=0.49$; $r_{14.25678}=0.35$; $r_{15.24678}=0.28$; $r_{16.24578}=0.18$; $r_{17.24568}=0.00$; $r_{18.24567}=0.15$.

Thus we may say that each test in the American Council battery contributes effectively to our knowledge of certain general factors underlying 4-year scholarship. High-school record samples these same general factors to some extent and adds unique factors which make such records equal in predictive importance to the 5-test scores combined. But the refined statistical procedure of determining the weight of each separate variable in a regression equation for predicting scholarship makes a negligible improvement in prediction over the

[6] Test percentile (variable X_3 of Table 4) has been omitted from the multiple correlation because it is merely the total of the five subtests (X_4, X_5, X_6, X_7, and X_8). Likewise first-term grade-point ratio (X_9) has been omitted because it is a part of the criterion (X_1).

simple procedure of averaging test percentile rank with high-school percentile rank. $R_{1.245678} = 0.65$ and $R_{1.(2+3)} = 0.62$.

A laborious search for additional measures of predictive significance has been rather disappointing.

1. The Inglis test of English vocabulary at the college level gives fair correlations with scholarship but adds almost nothing to prediction, because it is correlated $r = 0.4$ with high-school record and $r = 0.7$ with American Council test score ($N = 800$). Thus it furnishes little unique information about students and raises the multiple correlation of our personnel records with scholarship in the third decimal place only.

2. Division of test performance into linguistic and quantitative abilities in spite of the light it seems to throw on the intellectual make-up of individual students does not add appreciably to prediction.

3. We have tried out a homemade test of ability to take notes, with an added feature designed to indicate persistence. Neither measure gives any important new information about the potential scholarship of entering freshmen.

4. We have found the difference between test percentile rank and high-school percentile rank very revealing in individual cases, but these differences apparently do not represent any single trait in linear fashion, hence fail to improve general prediction.

5. Health records and data from the physical examination made at entrance give little or no indication of relative scholastic achievement, although they are certainly very useful in advising with individual students.

6. Measures of interest of the extrovert-introvert type and estimates of time spent in study show considerable promise as indications of scholarship, but we have been unwilling to use them systematically; because as soon as students realize that mere statements of their attitudes and habits are to be used administratively, the frankness of such statements becomes doubtful.

7. After discovering two boys in college whose elementary school preparation as measured by Stanford Achievement barely equalled that of the average seventh and eighth grader, respectively, we share the belief of Dr. Luella Pressey (5) that a representative sample of elementary-school tool skills might afford considerable predictive as well as diagnostic information about college scholarship. However, we have as yet developed no test for such a purpose.

Instead, we have extended our sampling of student performance into the college situation. Does scholastic performance during the first quarter in college mean anything with reference to 4-year scholarship? Advocates of freshmen week and sentimentalists in general picture the terrific adjustment problems of students away from home, lost in crowded classrooms, lonesome and bewildered by the details of new social and intellectual requirements as if performance under

such conditions were quite unrepresentative of real potentiality. But if 4-year college scholarship be considered an adequate criterion of scholastic potentiality, first quarter grades are on the contrary highly indicative ($r=0.78$ for these 331 students). Of course this correlation is to some extent spurious in that first quarter grades are averaged into the total with which they are correlated, but this is not spurious for our purpose since it is merely the earliest possible prediction of total scholarship which we are seeking. But the high correlation of first quarter grades with the average of test score and high-school record ($r=0.57$) prevents our getting as much unique information from these first quarter grade-point ratios as might be anticipated. Using these three measures, test percentile rank, high-school percentile rank, and first quarter grade-point ratio, the multiple correlation with 4-year college scholarship is $R_{1\cdot 23}=0.74$ for 120 men, and $R_{1\cdot 23}=0.83$ for 211 women, and $R_{1\cdot 23}=0.81$ for the 331 students who graduated four years after entrance. Thus these three personnel measures have at least 64 per cent commonality with the quality of 4-year college scholarship. Still more important, the reliability with which the common factors prevading all three personnel measures, and also underlying all-around scholarship, are sampled is sure to be much higher for the combination than for any of the measures singly. Hence injustice to individual students is greatly minimized by using such a combination.

This demonstration of the equivalence in meaning of college scholarship and personnel measures is not complete for two reasons. In the first place, it seldom pays to run every speck of potential energy through the power house. It is usually better to allow many streams whose "head" of potential energy is slight to flow off into other channels where they are really much more serviceable. Now by comparing the variance ($\sigma 2$) in all these personnel measures for the whole entering group with the variance in these measures for those who graduated, it is possible to estimate what the correlation of the measures with 4-year scholarship would have been if the total range of entering talent had continued or been allowed to continue under the selective processes represented by scholastic grades (4). Thus the estimated correlation of test scores alone, with 4-year scholarship if all students entering had remained that long, would be $r=0.62$ instead of $r=0.53$ within the restricted range of those who graduate. Similarly, the estimated correlation of our combined personnel measures with 4-year scholarship if all who entered remained, would be $r=0.87$ instead of $r=0.81$ within the restricted range of those graduating. In the second place, even for engineers estimated input and obtained output never balance exactly because there is always some error in measurement. So in estimating scholarship from personnel measures, both the criterion and the measurements upon which the estimates are based are by no means perfectly accurate. But with a reliability

coefficient of $r_1 I = 0.97$ for the pooled grades of the criterion and probably about $r_1 I = 0.95$ for the combined personnel measures, the estimated true correlation of potential and actual achievement would not rise greatly above the est. $r = 0.87$ found by allowing for restricted range.[7]

Yet, surely this makes it evident that early in a student's college career we can gage with surprising accuracy the excellence of his future scholastic achievement. Unmeasured factors, even of interest and effort, can not be highly important. Apparently such interest and effort factors as really function are included rather completely in preparatory school records and first quarter grades. Thus we have measures of potential scholarship which indicate very well what the actual achievement in knowledge, technical skill, and productive scholarship will be under the transforming influence of university instruction. To be sure this in no way disparages the necessity of continuous painstaking college instruction as superficial thinkers have sometimes fallaciously argued. One might as well argue the futility of the power house in transforming potential energy into kilowatts. It is only because of exacting instruction that potential scholarship forecasts so definitely future scholastic achievements. But whatever the nature of the general factors which underlie all-around college scholarship, we can be sure our test scores and other personnel measures indicate potentially these same general factors. If the pooled judgments of instructors with reference to scholarship have value, that same value applies to test scores and other personnel measures. Or if test scores be accepted as measures of intellect, then college instructors as a group recognize and value such ability very definitely.

The practical significance of these findings can be illustrated by an experiment in correlating estimated scholarship with that actually obtained at the university. If the average of test percentile rank and high-school percentile rank be combined with first quarter grades in the university, the composite correlates $r = 0.825$ with 4-year grade-point ratio for the 172 students who graduated four years after entrance in 1925. Using the same weightings for these three personnel measures, we estimated the 4-year grade point ratio of the 159 students in the 1926 entering class who graduated four years after entrance. We then correlated the estimated grades with the grade-point ratio actually obtained by these students during their college career. This correlation was $r = 0.806$ which at the same time verifies our computations of the previous year and demonstrates the essentially stable significance of our personnel measures from year to year. Finally, we have used the same measures and weightings to estimate the average scholarship of practically all those entering the university

[7] $r = \frac{0.87}{\sqrt{.97}\sqrt{.95}} = 0.91.$

as freshmen in 1925 who stayed as long as one quarter.[8] These estimated grade-point ratios were then correlated with the average points per hour actually earned for as long as the student remained or was allowed to remain. The result was $r = 0.86$. It will be remembered that the estimated correlation of these measures with 4-year scholarship in such an unrestricted range was $r = 0.87$. Again the consistency of our data and computations is gratifying.

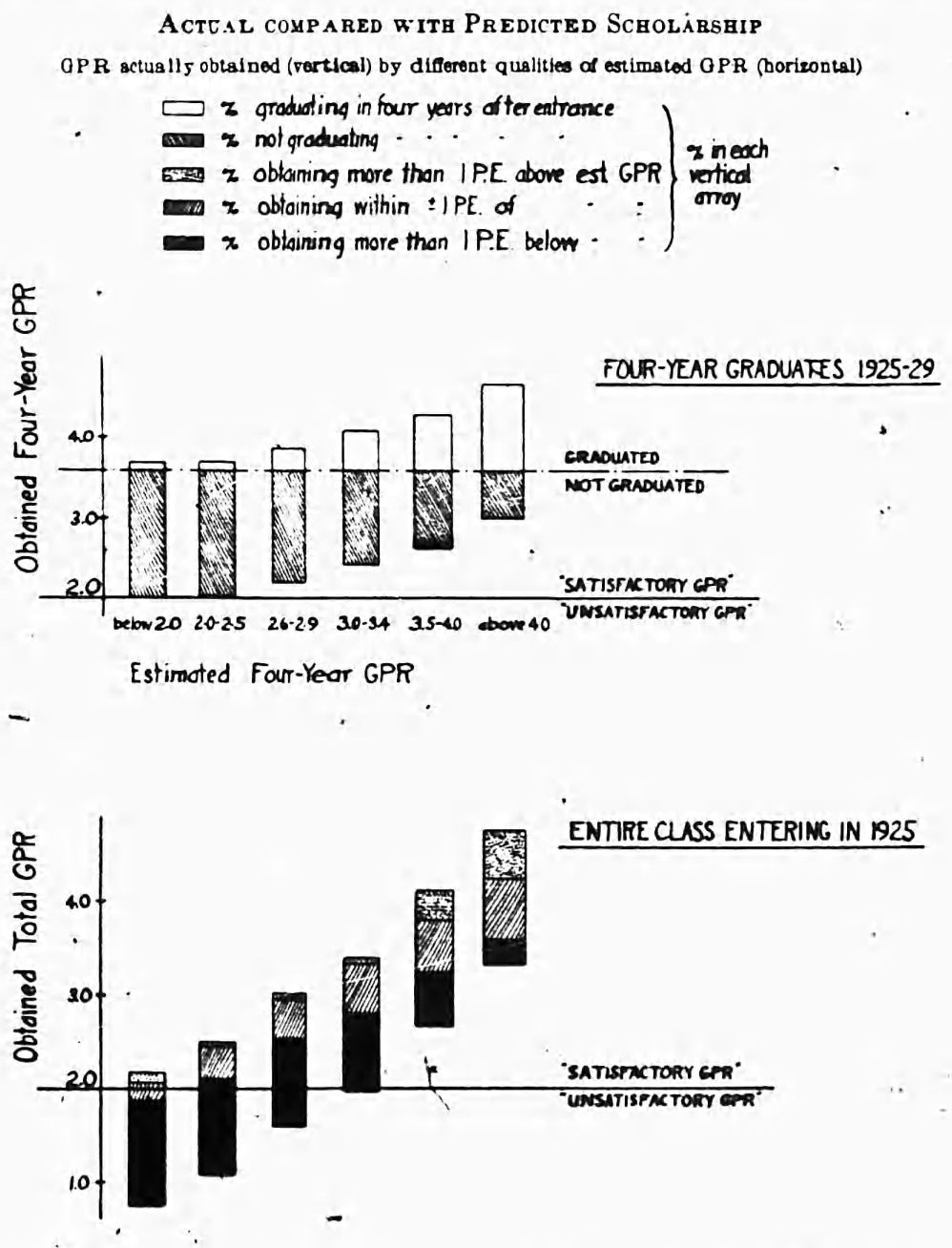
And now, briefly, what should be done about it? Well, so far as average quality of college scholarship is an adequate criterion, preparatory school records and psychological test scores are valid bases for advising very mediocre students not to attempt college work. Such an evaluation of college potentiality for each preparatory school student is feasible during his final high-school year. All that is needed is a uniform State-wide testing program and the preparatory school record. But if educational democracy requires an open door in State institutions for all high-school graduates, then certainly fairly rigorous elimination may begin at the end of the first quarter without any serious injustice and with great savings in time, money, and effort, not to say agony for all concerned. The following chart presents the evidence for this statement in graphic form.

Suppose in 1925 we had placed on probation at the end of the first quarter all students for whom our combined personnel measures predicted a 4-year grade-point ratio of 2.5 or less, and then disqualified all who failed to maintain a grade-point ratio of 2. Twenty per cent of the entering class would have gone on probation, but less than 3 per cent of our 1929 graduates, and only 11 per cent of those who graduated in five years would have done so. Presumably not a single graduate would have been disqualified. So far, students with grade-point ratios less than 2 simply do not graduate. Everything seems to indicate that the line between satisfactory and unsatisfactory scholarship at the University of Oregon might well be drawn at this point. Of course, students failing to earn grade-point ratios of 2 in any quarter should also go on probation and be disqualified if they do not improve, no matter how promising the scholarship indicated by their personnel records. A scholarship committee considering individual cases on their merits could easily furnish all the elasticity in administration which would be needed. The chief advantages of such a system would be:

1. When probation and disqualification are based on single quarter grades alone, as at present, the reliability of such judgments is quite low because even when the grades of a single quarter are averaged, the reliability of the resulting index is, in general, only $r_1 I = 0.73$,

[8] We utilized here data collected by the registrar, Dr. Pallett, for another purpose. Lack of other information about these students and withdrawal during the first quarter explains why 763 instead of the whole group were studied.

and much lower for possible combinations of departmental grades, to say nothing of the dubious reliability of a single-course grade, which often decides the whole matter. But when probation is based on potentiality estimated from personnel measures, including first-quarter grades, both the reliability and predictive significance of



such judgments in terms of 4-year scholarship are greatly improved. The same argument applies to basing disqualification on cumulative grade-point ratio instead of on the scholastic performance of a single quarter.

2. If probation were in part based on general ability test score, it would tend to stimulate each entering student to do his best on

the psychological examination, thus insuring more accurate measurement.

3. Likewise, if probationary status depended in part on preparatory school record or high-school content test score, there would soon be a salutary effect on the attitude of high-school students toward their preparatory school work.

4. Finally, scholastic leniency with freshmen is unjustifiable in view of the high correlation of first quarter grades with 4-year record. Such leniency encourages excuse making and refusal to face the facts so far as scholarship is concerned. Under the proposed more rigorous requirements students with real potentiality would be "put on their toes" at once instead of being confirmed in habits of preparatory school loafing. There is no better time to break with inefficient habits than when changing from accustomed surroundings and associates to new ones. As William James said in his famous essay, "a complete break is far better than a gradual one." For students without any scholastic future, an early recognition of that fact will go far to prevent a costly and tragic struggle against probability into which very mediocre students are lured by lenient scholastic requirements for freshmen. What we need is less sentiment and more psychology in freshmen week and personnel procedures generally.

BIBLIOGRAPHY

(1) Educational Record, April, 1926.
(2) Ellis, R. S. The correction of constant errors in college marks. School and Society, 24:432–436. 1926.
(3) Finkelstein, I. E. The marking system in theory and practice. Baltimore, Warwick & York, 1913. 88 p.
(4) Kelley, T. L. Statistical methods. New York, The Macmillan Co. 1923. Formula 182.
(5) National Society of College Teachers of Education. Yearbook XVIII, 1930. p. 167.
(6) Rugg, H. O. Teachers' marks and marking systems. Educational Administration and Supervision, 1:117–142. 1915.
(7) Spence, R. B. The improvement of college marking systems. Teachers College Contributions to Education, No. 252, 1927. 75 p.
(8) Thurstone, L. L. Psychological examination for college freshmen. Educational Record, v. 8, No. 1, January, 1927.
(9) Tryon, R. C. The interpretation of the correlation coefficient. Psychological Review, September, 1929, v. 36, No. 5, p. 419–445.
(10) Zerbe, J. L. Distribution of grades. Journal of Educational Psychology, 8:575–588. 1917.

A STUDY OF THE COLLEGE APTITUDE AND ABILITY OF HIGH-SCHOOL SENIORS

JOHN S. JORDAN[1]

I. INTRODUCTION

The survey reported herein originated in the desire to obtain certain specific information about the students coming to the Washington State Normal School at Ellensburg as compared with high-school seniors in general in the State, and as compared with those high-school seniors going on to other institutions of learning.

Some of the principals and superintendents of Washington have complained from time to time, perhaps rather uniquely, that our graduates are not always phenomenal teachers. We in turn have complained that the high-school seniors sent to us by these same superintendents are occasionally lacking in some of the academic virtues. This survey was suggested at a meeting of principals and superintendents of Yakima County, Wash., at which a representative of this school was present. The futility of the expression of personal opinions upon such matters was evident to many of those participating. The survey was, therefore, undertaken with the hope of supplying some objective evidence bearing upon the issues involved.

The writer wishes to express his appreciation of the fine courtesy and cooperation of the superintendents and high-school principals of Yakima County. Without this spirit, the study reported below would have been impossible. In many instances the giving up of half a school day for the testing program must have been inconvenient. The clerical labor necessary for the recording of marks and other data by administrators or their assistants was a burden which was cheerfully assumed by most of those involved. An attitude of this sort is a hopeful indication of future progress in educational investigations.

II. PURPOSES OF THE SURVEY

1. To supply information or data of value to superintendents and principals for use in the administration of their respective high schools.
2. To discover the range and status of college aptitude in a typical agricultural county of the State of Washington.

---

[1] John S. Jordan, head of the department of psychology, State Normal School, Ellensburg, Wash. B.A., University of Denver, 1916; M. A., Stanford University, 1923. He was formerly a member of the faculty of the Department of Psychology and Education of Colorado College.

3. To determine the relative status of students entering teacher-training institutions.

4. To determine the relative aptitude of the men as compared with the women students.

5. To determine the relation between high-school marks in several departments and aptitude test results.

6. To determine the relative values of high-school marks in the earlier years of high-school work, as compared with marks obtained in the last two years and the last year.

7. To discover the relative college-aptitude status of the seniors of small, medium, and large high schools.

8. To determine the relation between college aptitude and the degree of acceleration or retardation in the elementary school and the high school.

9. To determine the relationship, if any, between college aptitude and varying proportions of rural and town elementary schooling.

10. To discover the relationship between college aptitude, as measured by the tests used, and the occupation of the father.

11. To discover the relationship between the prediction index and the occupational intentions of the high-school seniors.

12. To determine the relationship between the prediction index and the educational intentions of the seniors, with separate analyses for different types of intended higher institutions, and for different sorts of training, such as liberal arts, engineering, teacher-training, etc.

13. To follow up, from the standpoint of the results obtained, the seniors who may enter upon their first-year course of preparation for teaching.

It is realized that many of the above-mentioned objectives are imperfectly accomplished in the study. This is due to many causes, among which are, smallness of sampling, inadequate methods of measurement, and limited means for securing data of a reliable sort.

III. PROCEDURE

1. The battery of tests, which is used at the normal school for all incoming students, was administered to the high-school seniors at the following schools of Yakima County: Cowiche, Granger, Lower Naches, Mabton, Moxee, Naches, Outlook, Selah, Sunnyside, Tieton, Toppenish, Wapato, Yakima, Zillah. The total number of students tested was 458. This group of tests has been chosen for use at the normal school, after considerable trial and experimentation, as meeting the needs of a teacher-training institution. It is probable that some other combination might be better for general college purposes, but it was believed that the advantage of comparing the performances of the high-school seniors with those of incoming normal-school students would more than offset any possible disadvantages. The tests used were as follows:

(*a*) The Detroit Advanced Intelligence Test is a variegated collection of performances representing a wide sampling of functions of a sort which most high-school students have had an opportunity to learn. The test is not based upon specific units of subject matter, but is quite general in nature, and is rather typical of the sort of instrument commonly called a "group-intelligence test." The Army Alpha was a pioneer test in this field. Such tests should not be thought of as direct measures of innate capacity, nor as a discrete measure of "pure intelligence." The object of its use was to obtain an all-around measure of college aptitude. Evidence from schools using such tests indicates that they are of some value in predicting college success as represented by academic marks.

(*b*) The new Stanford Arithmetic Test is a survey test including two subtests, one on the fundamental operations, and one for reasoning problems. The problems are representative of those commonly offered in the public schools in the grades through the ninth. All students in the Washington State Normal School, preparing for a teacher's diploma, who fall below the ninth grade norm, are required to take remedial work without credit until the deficiency is removed.

(*c*) The Iowa Comprehension Test contains selections of material similar to that found in many college textbooks in the fields of science, history, and literature. Fifteen questions are provided for each of the three selections. The test is intended to measure the ability to read understandingly such material at a fair rate of speed.

(*d*) The Purdue English Test covers certain fundamentals of English as needed in everyday life and particularly by teachers. The topics covered are punctuation, grammar, choice of words, literary information, spelling, vocabulary, and reading. Normal-school students falling below an empirically determined standard are required to take remedial work in English without credit.

2. The tests were administered, with one exception, by J. S. Jordan, of the psychology department, who is in charge of testing at the normal school. The tests take almost three hours to administer. The entire morning or afternoon session was consumed in each instance, with a 5-minute intermission at about the midway point. The standardized directions and time limits were adhered to absolutely.

3. The tests were scored by advanced normal-school students who were paid for their work, and who were under close supervision. Each test was scored independently by two different people. The standardized scoring directions were strictly adhered to.

4. A prediction index (P. I.) was computed for each student. The prediction index is a composite derived from the scores of the entire battery of tests. In arriving at this composite, the scores are weighted in proportion to the amount contributed by each test to

college marks. The weights were based on data from about 500 normal-school students. The contributions are determined by the multiple correlation method applied to the marks as the criterion and the test scores as the variables. Weightings are also included for differences in absolute size or score. The prediction index is a means of expressing by a single number, a prediction based upon standardized test results, as to a student's probable success in college work. The weighting of the prediction index is computed so as to yield an average of 100. The evidence is that the prediction index is more accurate than any single test score or any unweighted combination. The multiple correlation obtained between the prediction index and marks for the fall quarter of 1929 was 0.792 with a probable error of 0.017. This multiple correlation coefficient is the highest possible correlation between the composite of the test scores, each one weighted in optimum manner, and the dependent variable, namely, normal-school marks. The prediction index is the concrete representation of the best weighting plus the other adjustments referred to above.

5. The data concerning the occupational and educational intentions of the students, and other personal information were secured in most of the schools from a mimeographed blank, filled in by students under the direction of the examiner during the testing period. The blanks were not ready for use in the first schools tested. For these schools the blanks were sent to the superintendent with the request that they be completed. The high-school marks for four years were secured from the official records of each school. The test data are complete for 458 high-school seniors. The data are not complete for the other information due to the failure of a few superintendents or principals to make complete returns. But the deficiencies are small compared to the number for which information is complete, and we seem justified in assuming that all of the data represent fair samplings.

IV. RESULTS

1. Distributions of test scores and prediction indexes.

TABLE 1.—*Summary of results from 458 Yakima County high-school seniors combined, and 227 first-year normal-school students, fall 1929*

IOWA COMPREHENSION TEST

Measures	High-school seniors	Normal-school freshmen
1	2	3
Mean	26	26
Standard deviation	7.1	7.6
V (coefficient of variability)	27.3	29.2
Q (quintile)—1	43-31	42-33
2	30-26	32-28
3	25-23	27-25
4	22-20	24-20
5	19-00	19-10
Standard error of difference	.60	.....

TABLE 1.—*Summary of results from 458 Yakima County high-school seniors combined, and 227 first-year normal-school students, fall 1929*—Continued

DETROIT INTELLIGENCE TEST

Measures	High-school seniors	Normal-school freshmen
1	2	3
Mean	129	124
Standard deviation	31.7	31.5
V	24.6	25.4
Q—1	227–153	201–149
2	152–135	148–131
3	134–120	130–115
4	119–104	114–103
5	103–50	102–39
Standard error of difference	1.6	.....

NEW STANFORD ARITHMETIC TEST

Mean	104	103
Standard deviation	11	11.03
V	10.6	10.7
Q—1	125–114	124–114
2	113–110	113–108
3	109–103	107–102
4	102–95	101–95
5	94–65	94–70
Standard error of difference	.90	.....

PURDUE ENGLISH TEST

Mean	99	99
Standard deviation	17.4	15.8
V	17.6	16
Q—1	143–115	130–115
2	114–105	114–104
3	104–97	103–95
4	96–87	94–87
5	86–43	86–60
Standard error of difference	1.3	.....

PREDICTION INDEX

Mean	102	101
Standard deviation	19.3	20.2
V	19	20
Q—1	156–118	142–118
2	117–106	117–105
3	105–97	104–95
4	96–84	94–85
5	83–53	84–54
Standard error of difference	1.6	.....

It is to be noted that the means for the two groups in the Iowa Comprehension Test, the new Stanford Arithmetic Test, and Purdue English Test, and the prediction indexes are practically the same. The high-school seniors are slightly superior in the Detroit Advanced Intelligence Test, the difference being more than three times the standard error of difference. The standard error of difference is a means of indicating the significance of a difference between two measures. It expresses the probability of such a difference being erroneous

because of smallness of sampling. If the standard error of difference is equal to the difference between two measures, there is approximately one chance out of three that a complete sampling would show no difference. If the difference between two measures is three times the standard error of difference, the chances are 369 to 1 that there is a true difference. A 3 to 1 ratio between a difference and its standard error is considered by most statisticians to indicate practical certainty of the existence of such a difference. In illustration, the difference between the mean score of the high-school seniors in the Detroit Advanced Intelligence Test and the mean score of the normal school freshmen in the same test is 5 points. The standard error of difference is 1.6. The difference between the means, namely 5, is slightly more than three times the standard error of difference, which is 1.6. Therefore, the difference may be considered as almost certainly representing a true difference. On the other hand, the difference between the mean prediction indexes of the two groups is only 1 point. The standard error of difference is 1.6. Therefore, we can not say that a difference has been established. The highest average prediction index for any high school was 113. The lowest average was 81.

The variability of the normal school freshmen is very similar to that of the high-school seniors in all of the measures. The term variability refers to spread or dispersion within a group. High variability means a wide range of scores on either side of the mean. Low variability means a concentration of scores close to the mean. The standard deviation is considered to be the most reliable means of expressing the amount of variability. But the standard deviation is in terms of test units, so if a comparison is to be made between the standard deviations of two tests having different units, as is usually the case, the comparison is difficult to interpret. The coefficient of variability is in terms of the standard deviation and the mean. The mean may be thought of as the most representative score; therefore V, or the coefficient of variability, of any distribution may be compared with V of any other distribution, because all V's are rendered comparable through the equating influence of the means. Comparison of the V's reveals some interesting information. Both groups of students are most spread out in reading ability, as indicated in the results of the Iowa Comprehension Test, the variability being almost twice that of a normal probability curve. The spread is also marked for intelligence test scores. The dispersion is close to that of a normal distribution for the prediction index and English Test results. The results from the arithmetic tests show that the variation in this subject is very small, or in other words, that the students are more alike in arithmetic than in any other function measured in this survey.

2. Comparison of boys with girls.

TABLE 2.—*Comparison of high-school boys with high-school girls*

Test	Boys			Girls			
	Mean	Standard deviation	V	Mean	Standard deviation	V	Standard error of difference
1	2	3	4	5	6	7	8
Iowa Comprehension	26	7.32	28	26	7.18	27	0.68
Detroit Advanced Intelligence	132	29.82	22	123	27.99	22	2.70
New Stanford Arithmetic	108	10.00	9	101	6.64	6	.80
Purdue English	97	17.56	18	101	17.16	17	1.60
Prediction Index	102	19.92	19	102	19.02	18	1.80

According to Table 2, the high-school girls are slighly superior in English. The boys are significantly ahead in the intelligence test scores and even more superior in arithmetic. There is no sex difference in the reading test scores, nor in the prediction index. Compared to the total size of the scores, the differences are slight. Certainly these data give no basis for the opinion sometimes expressed by teachers that girls are brighter than boys.

In all of the measures except the intelligence test, the high-school boys are slightly more variable than the girls. The only measure for which this greater spread is marked is the arithmetic test.

Data secured over several quarters from normal school students indicate slightly greater variability among boys than among girls.

TABLE 3.—*Comparison of normal-school boys with normal-school girls*

Test	Mean score		Standard error of difference
	Boys	Girls	
1	2	3	4
Iowa Comprehension	25	26	1.1
Detroit Advanced Intelligence	116	128	4.0
New Stanford Arithmetic	102	103	1.3
Purdue English	96	100	2.5
Prediction Index	95	104	2.9

Table 3 indicates that the normal school girls are superior to normal school boys in all tests. The superiority is negligible in the Iowa Comprehension Test and in the arithmetic test, but is probably significant in the English test and quite marked in the intelligence test.

The situation is quite different from that found among the high-school seniors. Owing to the limited sampling interpretations must be made cautiously, but apparently the normal school girls beginning their first year are slightly superior to high-school senior girls, while normal school boys are below the average of high-school senior boys.

3. The relationship between size of high school and performance of seniors.

TABLE 4.—*Relative standings of small, medium, and large high schools*

Measures	Small high school	Medium high school	Yakima High School
1	2	3	4
IOWA COMPREHENSION TEST			
Mean	24	26	28
Standard deviation	7.2	7.6	6.7
Q—1	41–30	41–32	42–34
2	29–26	31–27	33–30
3	25–23	26–23	29–27
4	22–18	22–19	26–22
5	17–5	18–0	21–9
Standard error of difference between small and medium	.88		
Standard error of difference between medium and large	.77		
Standard error of difference between small and large	.84		
DETROIT INTELLIGENCE TEST			
Mean	122	126	137
Standard deviation	32.8	29.6	31.2
Q—1	227–148	203–151	227–161
2	147–130	149–133	160–144
3	129–114	132–120	143–130
4	112–98	119–101	129–113
5	97–50	100–71	111–58
Standard error of difference between small and medium	3.8		
Standard error of difference between medium and large	3.3		
Standard error of difference between small and large	3.9		
NEW STANFORD ARITHMETIC TEST			
Mean	102	105	105
Standard deviation	12.3	11.1	10.2
Q—1	119–112	122–114	125–115
2	111–107	113–110	114–111
3	106–100	109–104	110–106
4	99–92	103–96	105–98
5	91–65	95–71	97–69
Standard error of difference between small and medium	1.4		
Standard error of difference between medium and large	1.1		
Standard error of difference between small and large	1.4		
PURDUE ENGLISH TEST			
Mean	93	97	105
Standard deviation	18.8	17.6	16.5
Q—1	128–107	142–114	143–120
2	106–97	113–105	119–109
3	96–89	104–95	108–101
4	88–74	94–85	100–94
5	73–55	84–41	93–43
Standard error of difference between small and medium	2.2		
Standard error of difference between medium and large	1.8		
Standard error of difference between small and large	2.2		
PREDICTION INDEXES			
Mean	96	101	108
Standard deviation	19.8	19.2	18.7
Q—1	146–113	150–118	156–125
2	112–101	117–109	124–112
3	100–93	108–98	111–102
4	92–77	97–86	101–93
5	76–54	85–53	92–61
Standard error of difference between small and medium	2.3		
Standard error of difference between medium and large	2.0		
Standard error of difference between small and large	2.4		

Nine high schools were included in the small-school group. The average number of seniors per school in this group was 13. The range was from 10 to 17 seniors.

Four schools were classified as medium-sized. The average number of seniors was 44, with a range of 29 to 54.

One school fell in the large school category with 170 seniors taking the tests.

It will be noted that there is a direct relationship between size of score in each test and the size of the schools when classified in the manner referred to above. There are individual exceptions to this, however. The school, the seniors of which rank first in average score in each test and in the prediction index, is in the small-school group.

4. Relationship between marks and prediction index.

TABLE 5.—*Correlations between high-school marks and prediction indexes*

1	Number of cases	Coefficient of correlation with prediction index	Probable error of coefficient of correlation
	2	3	4
Marks for 4 years	339	0.367	0.032
Marks for second, third, and fourth years	343	.371	.032
Marks for third and fourth years	343	.394	.031
Marks for fourth year	342	.379	.031
English marks	343	.423	.030
Mathematics marks	333	.304	.034
Science marks	341	.312	.033
Foreign-language marks	274	.397	.037
Social-science marks	244	.315	.039
Commercial marks	223	.165	.044
Music and drawing marks	57	.142	.089
Home economics and industrial arts	155	.165	.052

The differences between correlations with the prediction index, of 4 years of high-school marks, the last 3 years, the last 2 years, and the last year are so slight as to be negligible. In other words the last 2 years of high-school work correlate at least as well with the battery of tests used as do the marks for all 4 years. As the composite of the tests has a proved prediction value for college marks, it would seem that much clerical labor might be saved if the higher institutions requested high-school principals to draw up transcripts of marks for the last 2 years only. Of course students would have to be certified as having met high-school graduation requirements, and as having taken certain prerequisites for entrance to specific courses.

It is interesting to note that marks in English and foreign language correlate significantly higher with the prediction index than do those for any of the other subjects. In fact, either one of them is at least as predictive as the composite of all marks for any number of years. Marks in commercial subjects, music and drawing, home economics, and industrial arts correlate so poorly as to be practically valueless for general prediction purposes.

TABLE 6.—*Correlations of normal-school marks, fall quarter, 1929, with test scores, prediction indexes, and high-school marks*

	Purdue English	New Stanford Arithmetic	Detroit Advanced Intelligence	Iowa Comprehension	Prediction index	High-school marks
1	2	3	4	5	6	7
Normal-school marks	0.55	0.47	0.58	0.61	0.79	0.61
Probable error	.02	.04	.03	.03	.02	.05

Correlation coefficients were computed between the test scores of normal-school students, fall quarter 1929, and their scholastic marks for the ensuing quarter. Also a coefficient of correlation was computed between mean marks and the composite or prediction index. These are presented in Table 6. A correlation coefficient is also included between normal-school marks and four years of high-school marks. It is to be noted that the Iowa Comprehension or reading test shows the closest relation with academic standing in the normal school, and that the arithmetic test shows the lowest agreement. The prediction index gives a significantly higher correspondence than any one of the tests.

It is surprising that the high-school marks do not show a higher correspondence with normal school marks than they do. The composite of four years of high-school marks show a definitely smaller correlation with normal school marks than does any one of the tests and a very much smaller agreement than does the prediction index. This may be partially accounted for by the wide variations found in the marking systems and the difficulty in equating them satisfactorily. Also, quite different standards of severity of marking obtain in different schools and for different teachers in the same school.

5. Acceleration and retardation and the prediction index.

TABLE 7.—*Relation of varying degrees of acceleration and retardation with test scores and prediction indexes*

Degree of acceleration or retardation	Mean scores				
	Iowa Comprehension	Detroit Intelligence	Stanford Arithmetic	Purdue English	Prediction indexes
1	2	3	4	5	6
ELEMENTARY SCHOOL					
Accelerated 3 years	30	170	111	103	126
Accelerated 2 years	26	142	104	103	106
Accelerated 1 year	29	143	109	107	112
Normal rate	26	128	104	99	101
Retarded 1 year	22	111	99	81	87
Retarded 2 years or more	21	125	99	84	89
HIGH SCHOOL					
Accelerated 1 year	30	149	113	115	104
Accelerated ½ year	25	126	105	97	100
Normal rate	26	131	105	100	103
Retarded ½ year	26	126	98	98	100
Retarded 1 year or more	23	117	103	91	94

Those pupils who were accelerated three years in the elementary school show a marked superiority in all tests and in the prediction index. The ones accelerated one or two years are superior in the prediction index but not in all of the tests.

The pupils who were retarded one or two years in elementary school are significantly below those who have progressed regularly, in all test scores and in the prediction index. Apparently there is little difference between those retarded one year and those retarded two years. The number of cases in each category is too small to give results of much worth.

The data on high-school irregularity of progress may be interpreted in a somewhat similar fashion.

6. Rural and town schooling compared.

TABLE 8.—*Relation of mean scores of varying proportions of rural and town schooling to test scores and prediction indexes*

	Number of cases	Iowa Comprehension	Detroit Advanced Intelligence	Stanford Arithmetic	Purdue English	Prediction index
1	2	3	4	5	6	7
All rural	65	24	124	104	93	97
Standard deviation all rural		6.8	32.4	11.3	17.8	20.1
¼ rural, ¾ town	8	24	118	106	96	97
½ rural, ½ town	29	25	126	106	98	101
¾ rural, ¼ town	18	25	131	105	101	101
¾ rural, ¾ town	33	25	123	100	100	99
¼ rural, ¼ town	18	26	131	107	100	103
All town	176	27	135	106	103	106
Standard deviation all town		7.2	29.6	10.1	15.7	18.3
Standard error of difference between all rural and all town		1.0	4.5	1.6	2.5	2.8

A rural school is arbitrarily defined as a 1 or 2 room school. A school having three or more rooms is defined as a town school. The results shown in Table 8 indicate that those receiving all or most of their elementary schooling in town schools test higher than those whose schooling has been entirely or mostly rural. The differences found between the intermediate proportions are inconclusive. These findings, of course, are not necessarily indicative of the superiority of town elementary schools. The native talent of children living in towns may be a factor.

7. The prediction index and occupation.

TABLE 9.—*Relation between occupations of fathers and prediction indexes of high-school seniors*

	Occupations of fathers					
	Profes-sional	Business	Salesmen	Clerical and skilled labor	Farmers	Laborers
1	2	3	4	5	6	7
Per cent of total group	7	16	5	16	53	7
Mean prediction index of children	112	110	115	103	99	109

In the above table the percentage of farmers is much greater than for the population of the entire State. This is to be expected, of course, in an agricultural valley such as comprises the inhabited portion of Yakima County.

The high-school seniors whose fathers are in the professions, in business, or who are salesmen make significantly higher scores than those whose fathers are clerical workers, skilled artisans, or farmers. The average performance of those whose fathers are unskilled laborers is distorted by the very high scores of two individuals. The total number of laborer's children is so small that these two cases have an unusual weight. There is much overlapping between the occupational groups. Therefore, occupation of father seems to be an unsound basis for the prediction of the scholastic performance of the offspring.

TABLE 10.—*Occupational intentions of seniors compared with their prediction indexes*

	Occupational choice of student					
	Profes-sional	Business	Salesman	Clerical and skilled labor	Farmer	Laborer
1	2	3	4	5	6	7
Per cent of group	45	7	1	40	7	0
Mean prediction index	109	102	111	99	97	.....

Almost half of the group expresses an intention of entering a profession. It is to be feared that some of them will be disappointed, as this is several times the percentage of people actually engaged in the professions. Only 7 per cent of the high-school seniors desire to be farmers. It seems regrettable that such a small percentage wishes to engage in the principal activity of Yakima County.

The number electing to become salesmen is so small, only 1 per cent, that it can not be accepted as a fair sampling as far as the prediction index is concerned. Omitting this group from consideration, the prediction indexes vary consistently in the same general manner as do the prediction indexes of the offspring of those included in the occupational groupings in Table 9. In brief, there is small but probably significant relationship between the prediction indexes and the occupational intentions of the high-school seniors of Yakima County.

8. Relationship between educational intentions and the prediction index.

TABLE 11.—*Test scores and prediction indexes in relation to intentions of continuing formal education*

	Per cent of total	Mean scores				
		Iowa Comprehension	Detroit Advanced Intelligence	Stanford Arithmetic	Purdue English	Prediction index
1	2	3	4	5	6	7
Students continuing with education	91	28	129	106	102	106
Students not going on to school	9	25	125	103	94	99

The average scores of students who plan on continuing their formal education beyond the high school are slightly higher in all of the tests and in the prediction index. The smallness of the superiority of those intending to attend college would indicate that aptitude for higher education is a small factor in influencing the decision of the pupil or of his parents. An untabulated analysis of the data shows that a large percentage of low-test-score people intend entering higher educational institutions. Many of these people will be eliminated at considerable expense to the State, and, perhaps more important, with attendant personal humiliation and sense of failure. There would seem to be a vital need for a program of guidance involving both the high schools and the higher institutions.

TABLE 12.—*Relationship between different types of higher educational intentions and test performance*

		Mean scores				
	Per cent	Iowa Comprehension	Detroit Advanced Intelligence	Stanford Arithmetic	Purdue English	Prediction index
1	2	3	4	5	6	7
Liberal arts	18	29	147	104	111	113
Business administration	7	28	140	110	106	110
Teaching (preparation other than normal school)	7	29	139	106	104	110
Engineering, science, etc	16	27	142	112	102	108
Professional, other than teaching	15	27	132	106	102	105
Teaching (normal school preparation)	7	26	130	106	98	102
Business college	20	25	122	102	98	99
Undecided as to course	6	26	120	105	95	99
College of agriculture	4	23	123	106	92	95
Standard error of difference between liberal arts and normal school		1.6	7.0	2.4	3.8	4.5
Standard error of difference between normal school and agricultural college		2.5	10.4	3.6	5.6	6.8

The above table should require very little interpretation. The people planning on attending a normal school made scores, which, on the average, approximated those of the entire group of high-school seniors, the prediction index being exactly the same.

V. INTERPRETATION AND SUMMARY

1. Mean scores of 227 first-quarter normal-school students are recorded in Table 1 (p. 53). In Table 6 (p. 59) correlations are given between the scores of 212 of these students and their first-quarter normal-school marks. When the letter marks are translated into an arbitrary numerical scale ranging from E-0 to A-10, the average is slightly in excess of 5 or C plus. The average marks of the respective prediction-index quintiles or fifths are given in Table 13.

TABLE 13.—*Showing for each prediction-index quintile group the mean normal-school mark*

Prediction index quintiles	Range of prediction indexes	Mean of normal-school marks	Standard deviation of marks
1	2	3	4
Q—1	142–118	6.3	1.64
2	117–105	5.7	
3	104–95	5.2	1.94
4	94–85	4.2	
5	84–54	3.4	1.82

Standard error of difference between means of normal-school marks of Q—1 and Q—3=0.37.
Standard error of difference between means of normal-school marks of Q—3 and Q—5=0.39.

One-half of those in the lowest prediction-index quintile do work averaging D or lower. D is passing but not satisfactory, and not more than one-fourth of a student's credits may be D.

Thirty-one people, or about 15 per cent of the group, earned a prediction index of 80 or below. The mean mark for the first quarter of 1929 for this group was 2.3, a trifle above D upon the above-mentioned scale. Only one person in this group received marks as high as the average for all first-year students. Twenty-six of the 31 received marks which were definitely unsatisfactory. These data fortified by data from previous years, would indicate that students with prediction indexes of less than 80 are very unlikely to do satisfactory normal-school work.

On the other hand no student in the upper fifth of the prediction-index distribution did definitely unsatisfactory work. Seven from 43 were below the average for the student body, but none of these was below an average of C.

The data presented in Table 13 and in the above statements should indicate, with a fair degree of reliability, the type of academic work to be expected from high-school graduates of varying prediction-index levels. This is particularly true of the upper and lower ranges.

2. Summary of results.—A brief summary of the results recorded in the tables and in the accompanying interpretative data follows. These conclusions are based upon the findings from 458 Yakima County high-school seniors and 227 first-year normal-school students. They are valid only in so far as these two groups are representative.

(*a*) Comparison of first-year normal-school students with high-school seniors.—(1) The average test performance of first-year normal-school students is about the same as that of Yakima County high-school seniors. (2) The highest high-school senior scores are above the highest normal-school scores. (3) The lowest scores are about the same for the two groups. (4) The normal school received a slightly higher proportion of the low-prediction people than of the very high prediction students. (5) Normal-school students and high-school seniors show similar wide variation in reading ability and narrow spread in arithmetic.

(*b*) Sex differences.—(1) Sex differences in test performances are small or negligible for high-school seniors. (2) Normal-school girls are slightly ahead of normal-school boys in two tests, and significantly ahead in the remaining two tests. The girls have an advantage of 9 points in prediction index, a difference which is statistically significant.

(*c*) Size of school as a factor.—There seems to be a tendency for the performance of high-school seniors to be somewhat in proportion to the size of the school. While this is true for the averages of the different size classifications, there are individual exceptions.

(*d*) Reliability of tests in prediction.—(1) The high-school marks for the last two years, or for the last year, correlate as well, if not better, with the prediction index than do the marks for three or four years. (2) English and foreign language marks correlate better with the prediction index than all of the marks for any number of years. They also show significantly better correlations than any other subject groupings. (3) Commercial marks, marks in music and drawing, and those in home economics and industrial arts show negligible correlation with the prediction index. (4) Any one of the tests used has shown better correlation with normal-school marks than the average of four years of high-school marks. (5) The prediction index has a very much higher prediction value than any one of the tests or high-school marks.

(*e*) Acceleration and retardation.—(1) There is a definite tendency for those high-school seniors who have been accelerated in either elementary school or high school to obtain higher prediction indexes than the average. (2) There is a tendency for those seniors who have been retarded in either elementary school or in high school to earn lower prediction indexes than the average.

(*f*) Rural-urban factor in elementary schooling.—There is a slight tendency for those having all or most of their elementary schooling in town schools to receive higher prediction indexes than those having all or most of their elementary schooling in rural schools.

(*g*) Occupation of parents.—There is a tendency for the children of professional men and those engaging in business to obtain higher prediction indexes than the children of artisans or farmers.

(*h*) Vocational plans.—(1) There is a tendency for high-school seniors who plan on entering a profession or business to have higher prediction indices than those who expect to engage in clerical work, skilled labor, or farming. (2) The percentage of seniors who are anticipating a white-collar job is probably much greater than can be accommodated. (3) Apparently farming needs to be made more attractive to high-school graduates. (4) More than 90 per cent of the Yakima County high-school seniors expect to continue with formal education. (5) There is a slight but probably significant relation between the intention of going on to college and aptitude for college work. (6) A considerable number of seniors expect to expose themselves to higher academic training who could probably profit more from some other type of training. (7) Table 12 (p. 63), shows the average scores made by those expecting to enter different types of higher education.

In interpreting the above correlations and apparent tendencies, we must be cautious in assigning causal relations. The fact of correlation does not prove causal sequence. The many factors involved are probably interlinked with each other in a complex interdependent fashion. Any factor may be as logically considered a dependent variable as an independent variable.

In conclusion, the writer realizes the limitations of this study. It is hoped that the survey may be charitably interpreted as an attempt to illuminate, however faintly, some of the many specific problems of the Ellensburg Normal School. Quite different results might have appeared if the seniors of large city high schools had been included. Such a survey carried on in other States might yield very different returns. The usefulness of the battery of tests employed would almost surely vary for other types of higher institutions. It is not assumed that any of the findings should be generalized without much more corroboration than is available.

REMEDIAL READING INSTRUCTION AS A PHASE OF PERSONNEL WORK IN HIGHER EDUCATION

FRANK W. PARR [1]

I. IMPORTANCE OF THE PROBLEM

The purpose of this paper is to give a brief discussion of remedial reading instruction as a phase of personnel work in higher education. It is needless to say that one can not do justice to such a comprehensive and important topic as this in the short time allotted. However, I shall attempt as best I can to suggest in this paper the need for, and the nature and extent of, a procedure that might be used, and the effects of a program in remedial reading on the college level.

That poor reading ability is a distinct handicap to college students has been pointed out by such authorities as Morrison, Book, the Presseys, Remmers, Lemon, and others. In discussing the diagnosis of pupil difficulties, Morrison says that "cases are occasionally found in which pupils progress incredibly with very slender reading ability—a very considerable number of pupils find their way into high school and even into the college without the reading adaptation. They can get the meaning of the printed page, but they do so laboriously by a process of deciphering. In effect they are usually slow students, and when they reach the subjects which require assimilation by extensive reading they become problem cases. They can not study effectively subjects which require extensive reading because they can not reflect upon the meaning as they read" (2).[2] Lemon found that practically every member of his group of problem cases at the University of Iowa had a marked deficiency in reading, as did Remmers with his group at Purdue University. Book, working with freshmen at Indiana University who were unsuccessful in their university work, said that he found that these students were very deficient in their ability to read and had to be given special help in learning to read more effectively before they could succeed with their academic work. He says that his experience with these students "clearly showed that the difficulties which they were encountering were chiefly due to their inability to read, and to wrong methods of work" (1). The writer a few years ago made a study of the poor readers among the freshmen who entered the University of Iowa in the fall of that year. He found that 63 per cent of the 350 poor readers received scholastic delinquency reports at the midsemester of their first semester's work, with an average of 5.7 hours work delinquent per student. Forty-

[1] F. W. Parr, professor of secondary education, Oregon State College. B. S., University of Illinois, 1925; M. A., University of Iowa, 1926, Ph. D., 1929. Publications: With E. R. Isvik, "Handwriting in the High School," *School Review*, 35:776–779, December, 1927; "A Remedial Program for the Inefficient Silent Reader in College," *Phi Delta Kappan*, 12:58, August, 1929; with C. L. Nemzek, "What Becomes of the Inefficient Silent Reader in College," *Peabody Journal of Education*, 7:299–303, March, 1930; "The Extent of Remedial Work in State Universities in the United States," *School and Society*, 31:547–548, April, 1930; "How Do College Students Prepare an Assignment," *School and Society*, 31:712–713, May, 1930; "Teaching College Students How to Read," *Journal of Higher Education*, 2:325–331, June, 1931.

[2] Numbers in parentheses refer to "Bibliography," p. 71.

nine per cent of the grades received by this group at the end of the first semester were below C. By the beginning of the second semester 110, or 32 per cent of this group had been eliminated from college.

The importance of the situation is well stated by Schultz and Miller who report a reading investigation which was carried on at Christian College at Columbia, Mo. "The enrollment of first-year college students suffers approximately a 40 per cent mortality during each year. Some educators attribute the failure of many of these students to poor reading ability. All of the studies which have had to do with reading at the college level reveal the fact that students show a tremendous variation in ability to read. In extreme cases certain individuals have been found to read nineteen times as effectively as others in the same class. Since a college education must come largely through the medium of reading, any scheme which will improve reading with a reasonable expenditure of time and effort can certainly be justified on the basis of aggregate benefits which will accrue throughout a 4-year college course" (4).

II. THE EXTENT AND NATURE OF REMEDIAL READING INSTRUCTION IN COLLEGES THROUGHOUT THE UNITED STATES

Anyone interested in the remedial phase of reading would be impressed in examining our educational journals with the number of studies reported which have been carried on in the elementary school. He would be equally impressed with the paucity of reports pertaining to the upper levels of the school system. Two possible explanations might be advanced to throw light on this situation. In the first place, it may be that educators assume that students of college age have an adequate mastery of the reading process, and therefore need no further training along this line. A number of experiments have been carried on to prove that this assumption is quite erroneous. A more plausible explanation for the paucity of material on the college level is that those men who have been carrying on such work have been negligent in reporting it. That is, it is reasonable to assume that only a small per cent of the studies in any field of endeavor is reported in our educational journals.

In order to get more complete information on the extent and nature of remedial reading instruction in this country the writer two years ago sent a letter to every State university in the United States. This letter, which was in the form of a questionnaire, was addressed to the dean of the college of education at each institution. The following is a summary of the data received from the 40 schools that returned the questionnaire. (1) Only 9 schools made any attempt to discover the poor readers among their freshmen. (2) Only 7 of these schools had a plan for assisting the poor readers. (3) When remedial work was offered it was usually under the supervision of the college of education, although the psychology department assisted in the work

at 3 of the schools. (4) The remedial instruction when offered is a phase of a "How to Study" course. Seven of the 9 schools offered it in this manner. (5) Only 4 schools made the remedial work compulsory for those in the freshman class who were in need of such instruction. (6) Only 4 schools gave college credit for the remedial instruction. (7) There was no standard practice as to the length of time devoted to the remedial work or to the frequency of class meetings. The range in the length of time given was from 2 weeks to 36 weeks, and the number of meetings held ranged from one every two weeks to 3 meetings per week. (8) Five schools reported that they used a syllabus or workbook in connection with the remedial work. (9) In reply to the question, "Do you have any evidence that this work improves the reading ability of the students?" Only 5 schools replied in the affirmative. Five schools also claimed that they had evidence to show that the students did better college work in general as a result of their improved reading ability. (10) A number of the schools described briefly the nature of the remedial program. Some of these descriptions were: "Course in psychology of reading," "Only locate the trouble," "Mostly throat relaxation and increased eye span," "Merely diagnose reading comprehension of freshman—no remedial treatment."

That a great deal of interest is being manifested in this problem of remedial training in reading on the college level is indicated by the fact that deans of 16 schools where no remedial program was provided made comments on their questionnaires expressing keen interest in, and approval of such work.

III. SUGGESTED PROCEDURE FOR CARRYING ON A REMEDIAL READING PROGRAM ON THE COLLEGE LEVEL

In planning a remedial program for the poor reader in college one should follow some well-defined procedure, which would probably incorporate the following steps: (1) A complete case history of each student. This should give information concerning the student's school history, physical record, reading habits, emotional characteristics, and study habits. (2) An adequate diagnosis of disabilities of each student. This will necessitate the use of two principal media—viz, observation and testing. Through observation we must note the frequency and nature of the eye-movements, vocalization or lip reading, finger pointing, visual defects, etc. By means of tests we may get information concerning the general mental ability of the student, which is a very essential type of information for any remedial program, his general reading ability, and his specific reading abilities. A good diagnostic test should be used which will point out specifically the deficiencies in the various skills in reading (e. g., Iowa silent reading test, University of Minnesota reading test). (3) A remedial program based upon the analysis of deficiencies should be set up for each

student. Remedial drill exercises should be provided to care for each deficiency. It is needless to say that the remedial training for each student should be designed to fit his needs, and since no two students present exactly the same combination of deficiencies, the remedial instruction should be given individually if at all possible. It is well to test each student at least once during the course of the remedial instruction, and again at the end of the program so that both the student and the teacher may know just how much progress has been made. The more highly the student is motivated during the remedial program, other things being equal, the better will the results be.

IV. THE EFFECTS OF THE REMEDIAL PROGRAM

Published reports of studies which have been carried on in this field seem to indicate that remedial instruction in reading on the college level may be extremely beneficial to the student. Reference will be made at this time to a few of the outstanding programs that have come to the writer's attention.

Probably the most comprehensive program to be reported in remedial reading instruction is that which is under the direction of Drs. L. C. and S. L. Pressey at Ohio State University. A reading test is given to all freshmen entering that school, and all of the poor readers are required to attend a remedial class until their reading deficiencies are removed. Since this work has been carried on for a number of years the Presseys now have some evidence to show the effect of the instruction. In a controlled experiment, a group of 422 poor readers at Ohio State University was paired with a similar group which was not given remedial instruction. The experimental group far excelled the control group both in improvement in reading and in scholarship. In commenting on her study, Mrs. Pressey says, "It seems quite evident from this investigation that it is possible to train students to read effectively and that such training is more likely than not to transfer to the preparation of lessons and to general understanding of college work" (3).

Book working with 54 students at Indiana University reports that these students increased their reading efficiency on the average 102 per cent. Some of his group showed improvements as high as 250 per cent. The ability of the group to master a standardized assignment had likewise increased from 60 to 97 per cent.

Remmers and Stalnaker at Purdue University gave remedial training to 7 freshmen students who scored in the lowest quartile on the American Council Psychological Examination. The results show an average gain of 24.6 per cent in rate of reading and a similar gain in comprehension.

In an unpublished study Schultz and Miller describe an experiment carried on at Christian College, Columbia, Mo. Gains in reading comprehension for a group of 27 poor readers ranged from 0 to 114 per cent.

In the June, 1931, issue of the Journal of Higher Education, the writer describes an experiment which he conducted with 20 students at the University of Iowa, and shows the effects of a well-organized program of remedial instruction in reading. Gains in both comprehension and rate of reading were made by each of the 20 students; some of the gains being well over the 100 per cent mark. Not only did these students improve materially in reading ability, but they also showed gains in scholarship records. With the exception of four students, the members of the remedial class made gains in scholarship average. During the semester in which the remedial instruction was given, nine, or 45 per cent, of these students earned their highest grade point averages for any single semester in college. A follow-up study of the 16 students who were enrolled in the university the year following the remedial instruction showed that these students continued to improve in scholarship. Eighty-two per cent of the group earned scholarship averages which equalled or excelled those for the previous year. Six of the nine students who made record averages during the period of the remedial instruction earned even higher averages for the following year. Three other members of the remedial class also made record averages for the year following the reading instruction.

While the experiment just cited was carried on with upperclassmen, it is reasonable to assume that comparable results would have been obtained with an underclass group.

It is interesting to note that most of the investigators who have worked in this field agree that students profit by this remedial instruction in proportion to their mental ability. It probably does not pay to spend time and money on the subnormal group, for the improvement does not justify the expenditure.

In conclusion, the writer has tried to show (1) that there is a need for providing some type of remedial instruction for the poor readers who come, and probably will continue to come, to our colleges; (2) that there are but few schools throughout the country that give any attention at all to their poor readers; (3) that one must have a well-planned program for carrying on such remedial instruction; and (4) that the results obtained from studies of remedial reading seem to warrant the attention which is being given to this important phase of educational work.

BIBLIOGRAPHY

(1) Book, W. F. How to Succeed in College. Baltimore, Warwick & York, 1927. p. 96.

(2) Morrison, H. C. The Practice of Teaching in the Secondary School. Chicago, The University of Chicago Press, 1926. p. 620.

(3) Ohio College Association, Bulletin No. 55.

(4) Schultz, Esther *and* Miller, James C. Unpublished report on a reading investigation at Christian College, Columbia, Mo.

THE PREDICTION OF SUCCESS IN ENGLISH COMPOSITION

L. KENNETH SHUMAKER [1]

The prediction of success in English composition is a problem which has long been a serious concern of teachers of English. The multitude of factors which enter into the problem and the varying importance attached to these factors by independent workers in the field have served only to create confusion. It is our purpose to show to what extent certain research carried on in the English bureau of the University of Oregon has served to define and limit the different factors of the problem and to solve some of its difficulties.

Most composition teachers will agree that they are interested in teaching students to write organic English. They begin to differ immediately as soon as various contents of courses, methods of instruction, and tests for measuring the results of instruction are considered. Let any group of English teachers attempt as innocent a task as setting up "minimum essentials" and the troubles become instantly apparent.

In order to accomplish anything, it becomes imperative that terms and objectives be defined with greatest clarity. The universe of discourse produced thus arbitrarily, gives a known point of departure from which measurable progress may be determined and calibrated. In the English bureau, our objective was stated in this manner: Let us demand that no student be admitted to any college class in English composition until he can write an organic sentence. An organic sentence means a group of related, arbitrary symbols (words) which communicate a complete thought. The presupposition is allowed that thinking takes place within the mind of any given individual by means of the proper combination of ideas; and that the correlation between the effectiveness of the thinking process and the representation of it by means of words as arbitrary, common symbols of knowledge, accepted by two or more individuals, differ among individuals in unknown ratio. The tendency is that the clearest thinking is easiest to express in words, and that the clearest thinker tends to have at his command the largest stock of words in which to represent his ideas. There is no reason to presuppose that an individual who is inarticulate in words may not be preeminently articulate in music, painting, sculpture, or even mathematical symbols. It is also taken for granted that not only must a certain word be the symbol for one

[1] L. Kenneth Shumaker, supervisor of the English bureau, University of Oregon. B. A., University of Iowa, 1922.

idea in a given context, but that the arrangement of the words must follow a definite pattern as thoroughly agreed upon as the definition of the word itself within the pattern.

There are many ways in which to acquire the knowledge of words and word patterns, but the most commonly accepted methods are (*a*) through a study of rules (formal grammar) which state concisely, observed behavior of word patterns (sentence structure); or (*b*) through habitual reading, hearing, writing, and speaking in which symbols and patterns become habitual in the individual without his awareness of any names of descriptions of these habits. The teacher in elementary and secondary schools attempts to employ the first of these methods in the recognized "language" courses of the public schools. The everyday contacts of the individual from the time he first learns to say "papa" and "mamma" tend to fix habits which may not accord in any way with the formal rules taught in the classroom. There is every reason to recognize that the second method has the deepest and most lasting impression upon the mind of the individual because (*a*) to appreciate the significance of formal rules requires a certain ability to think abstractly, and (*b*) to acquire the habits of companions with whom an individual is continually thrown is easier than to learn to obey the arbitrary dicta laid down in 1 short hour of the 24, two, three, or five times a week during three-fourths of the year.

With these presuppositions in mind, therefore, it seems more logical to attempt to determine, first, the student's language sense, or general aptitude for language; and, second, to diagnose his specific difficulties with language before attempting to instruct him in college classes in English. Expediency also renders highly desirable the employment of tests which are as objective as possible for measuring aptitude and for diagnosing difficulties. There is a well-known tendency on the part of humanly frail readers to take into account too many connotations implied in the contexts of manuscripts in attempting to evaluate the pure denotations evidently expressed. To put it another way, there is a tendency to give the benefit of the doubt to any manuscript which contains cleverness, humor, or penetration, even though a comma may be misplaced or a modifier may be out of its accepted bounds. Such an error may be due to the student's inadvertence. Since the composition teacher is primarily interested in the excellence of the mechanics of language, particularly in the elementary courses, there is also the necessity of giving high value to the manuscript which is free from language errors, but which contains a slight modicum of cogence.

Reflection upon the material thus roughly indicated determined the steps which led to the construction of the objective aptitude test now in use by the English bureau. This test consists of four parts, each of which contains 100 possibilities of success, and each of which

is intended to measure a single, indispensable phase of language aptitude.

Part I is devoted to finding out whether or not the student can put the right word in the right place and whether or not he can spell that word correctly, if he does know it. It is quite true that there is no correlation recognized between a pure ability-to-spell and language sense, yet it is most desirable that words not only have but one definition in a specific context, but that they have a standardized orthography. To the extent that the student is able to determine the right word for the right place in this part of the test, we are measuring pure language aptitude; and to the extent that the student spells the word correctly, we diagnose his ability to spell.

The instructions for Part I are as follows:

> Fill each blank with a word according to the sense suggested. Make each word fit as accurately as possible.
>
> EXAMPLE: Thanksgiving comes the last Thursday in N_____.
> (Name of month.)

It will be observed that the stress is laid upon the aptitude aspect of this test and that there are no elements of confusion offered in the diagnostic aspect of the test. Most written spelling tests either spell the word correctly and incorrectly and ask the student to make a choice of the correct spelling, spell some words correctly and others incorrectly and ask the student to write the correct spelling of any incorrectly spelled words in a convenient blank, or ask the student to supply a few troublesome letters in words which are almost entirely spelled out in contexts or in columns. In any case, the tendency of all these tests is to call especial attention to difficulties and to give the least chance for common memory habits to assert themselves.

The words in Part I are arranged in order of difficulty. Those at the first of the test are implied most obviously and are most easy to spell. The very fact that the first part of the test is extremely easy is an encouragement to the student to proceed with expectation of success.

The predictive power of Part I is approximately double that of any other part of the test for the purpose of determining language aptitude.

Part II attempts to measure the student's aptitude for the use of correct idiom. It might, perhaps, be more accurately called a *usage* test. The instructions at the first of it are as follows:

> In the following composition 100 words or phrases have been underlined. Some of these are incorrect and some of them are correct. Draw a line neatly and clearly through the incorrect words or phrases. Make no other marks upon the paper.
>
> EXAMPLE: He ~~aint~~ going home.

The words and phrases selected have been chosen from evaluated test material used over a period of years and are arranged as nearly

as possible in the order of difficulty. They also appear in a normal setting as part of a context. Considerable difficulty was encountered in building a simple, easy-flowing composition which conformed to the necessary specifications. The composition had to be organic and natural, and it had to contain the 60 incorrect usages and 40 correct ones in order of difficulty as nearly as possible. The predictive power of Part II is approximately one-half that of Part I.

Part III is called *Punctuation*. It is divided into three sections, with instructions as follows:

SECTION 1. In the following composition 20 blanks occur. A period is the appropriate mark of punctuation which should be placed in *some* (not all) of these blanks. Place an × in the blanks where you think a period belongs.

EXAMPLE: Come to the house ___ at 4 o'clock × we shall have tea then ×

SEC. 2. In the following composition 20 blanks occur. A comma should be placed in some of them, a semicolon should be placed in others, and some should be left blank.

EXAMPLE: The red , blue , and yellow of the sky blended beautifully ; the sun seemed to be the center ___ of a great ___ glowing vortex.

SEC. 3. In the following passage you will find blanks in which certain marks of punctuation should be inserted. Sometimes several marks should be put into the same blank. Notice carefully the different single choices which may also occur in combinations:

1. Leave the blank open if no punctuation is required.
2. Use × to signify a period.
3. Use a comma according to the rules for the comma.
4. Use opening and closing quotation marks (" ") around direct quotations.
5. Use a question mark after questions.
6. Use ¶ to indicate a new paragraph.

You are not to change capitalization or make any other marks upon your paper.

EXAMPLES: " perhaps it is the very simplicity of the thing ___ which puts you at fault ," ___ said my friend × " what nonsense you do talk ! " ___ replied the inspector , laughing ___ heartily ×

The punctuation tested for in this part may be called "functional" in the purest sense of that term. None of the arbitrary punctuation marks used according to convention alone appears. A student may learn by rote such conventions as placing the period after abbreviations, the comma between the number of the day of the month and the number of the year, in the same way in which he learns to spell a word correctly, but he must have a certain aptitude for language before he knows that a period comes at the end of a declarative sentence or that a comma is placed after an introductory adverbial element. The predictive power of this part of the test lies between that of Part I and Part II.

Part III was made in sections in order to get some kind of "order of difficulty." Section 1, it is observed from the instructions just quoted, demands the use of end punctuation only. The correct use

of the period in this section means that the student has sufficient language aptitude to know when a complete thought has been expressed and another thought begins. In section 2 every effort was made to present undebatable uses of comma or semicolon. In section 3 a passage from a standard edition of one of Poe's tales was taken almost verbatim, because the greater our variety of choice in punctuation becomes, the more complex the context in which it appears, the greater is our tendency to differ with each other about the loci of marks of organic punctuation. The English bureau wished to be absolved from as much argument as possible.

Part IV is called *Grammar*. The instructions at the first of this part are:

> In the following composition 100 words or phrases have been underlined. Some of these words and phrases are correct and some of them are incorrect.
>
> Draw a line through those words or phrases which you believe to be incorrect.
>
> Do not try to improve upon the composition. Make no other marks upon the paper.
>
> EXAMPLE: There ~~is~~ four men in the room which adjoins the library.

It will be observed from these instructions that Part IV follows the technique employed throughout this test: The presentation of every item in as natural a contextual setting as possible, so that the student will have every opportunity to make maximum use of his language habits, whether or not he knows a single formal rule of grammar, because it is the purpose of this test to measure the degree of refinement of language habits, or language aptitude. This part of the test was lowest in predictive value and was arbitrarily rated at unity.

This completes a recapitulation of the theory upon which the objective aptitude test has been constructed, and a description of the physical form in which it appeared. No matter how long the problem of predicting success in English composition is pondered in theory, and no matter how beautiful the solution appears on paper, the actual use of the test is the ultimate answer to the question: Has the instrument practical value? The next step was to give the test.

The statistical department of the university was called upon to give its assistance at this point, and the research was placed under the direct supervision of Ralph Leighton. The reliability of the test was found to be surprisingly high. The only flaw seemed to be that there was no criterion with which to compare an apparently effective test.

Most composition teachers base their estimates of student achievement upon their rating of themes written by the student, but the variation in these ratings by different instructors who read the same themes—and even by the same instructor when he reads the same theme more than once—is notorious. Several reputable composition

scales were investigated and rejected because they did not afford sufficient control of the factors involved in the problem of obtaining highly reliable judgments. Research upon the evaluation of essay type examinations indicated that a sufficient number of judgments tended to correct the errors in each other, if these judgments were averaged in order to get a general merit score to represent the value of student work. We therefore devised a score which will be explained below, to serve as the means of gaining a reliable estimate of themes which would be written by a group of students at as nearly as possible the same time, and under the same conditions, as these students would take the objective test. The evaluation of these themes would become the criterion with which we should compare the objective test.

In evaluating each theme, the following equation was used:

$$GM = \frac{1F + 2G + 2R + 5C}{10}$$

It is explained in this manner:

GM = general merit.

1F = 1 weight given to a score of 100 points for a paper perfect in physical form. In estimating form, only the appearance of the paper is scored.

2G = 2 weights given to a score of 100 points for a paper perfect in all the mechanics of idiom, grammar, and correct sentence structure.

2R = 2 weights given to a score of 100 points for a paper having the most artistic and skillful use of word choice, sentence structure, and rhetorical excellence.

5C = 5 weights given to a score of 100 points for a paper presenting the best thought content and most logical organization of ideas.

The denominator 10 is the sum of the weights.

The use of this equation resulted in producing a criterion of extremely high reliability. The immediate explanation of this high reliability seems to be that the last three members of the numerator absorbed any widely differing points of view upon the different elements judged, so that the general merit of the paper was not unduly affected by any reader failing to include all of the possible merits of the paper in the final score.

Late in the spring term of 1929–30 a group of 291 high-school seniors in Eugene High School and the University High School was given the objective test and asked to write a theme upon a subject presented to them.

The objective test was carefully and accurately scored by trained clerical assistants, and the criterion was scored by two Portland high-school teachers of excellent training and experience. Miss Geraldine Cartmell and Mrs. Katherine Dilió.

The reliability of the objective test was found to be 0.93; that of the criterion, 0.88; the coefficient of correlation between the test and the criterion, 0.50. This correlation was three times as great

as the correlation between the criterion and the next highest correlating objective test among three others used in the experiment. The aptitude test was later found to correlate 0.67 with the psychological placement test.

Although the test was statistically evaluated and pronounced sound in technique, capable of administration, and highly reliable and reasonably valid against a reliable criterion, through this preliminary manipulation, the question "Has it practical value?" was not yet adequately answered. The third step was to use the test with the entering freshman class of the fall of 1930. Eight hundred and fifty-eight students were grouped in percentiles upon the basis of scores made in the test under discussion. We had the percentile rating for each student according to his high-school record and also according to his record on the psychological placement test. This gave us courage to make an arbitrary division between the seventeenth and eighteenth percentiles of the English aptitude test, because we could use these other two ratings in addition to actual class contacts to correct any injustices which might be done. We therefore arbitrarily assigned 158 students who were below the eighteenth percentile to our course in English A, designed for the correction of faults in technical English.

A study of the curve made by plotting the percentiles above the seventeenth would indicate that we could conveniently divide the group which was not assigned to English A into three parts. Those in the highest "third" should probably be placed in accelerated sections; those in the middle "third" should be in normal sections; those in the lowest "third" should be in retarded sections. The organization of our curriculum precludes a course in English composition in the freshman year, hence it is impossible to give any data which might support the validity of our sectioning students above the eighteenth percentile of our English aptitude test. We have the data upon the lowest group only.

Let us review briefly the known points with which we had to work in attempting to predict results in this remedial course in technical English: (*a*) We had more statistical information than is given in this report. This additional information is available in the statistical department of the school of education, but it is omitted here for practical reasons. (*b*) We had the psychological placement test percentile, the high-school record percentile, and the English aptitude test percentile for each student. (*c*) We knew from investigation that the practice factor in the aptitude test was practically nil.

Our administrative provisions for conducting English A demand that we divide our subfreshman group into thirds and that we give remedial instruction to each third for one full term—more or less—

at the discretion of the instructor. Of students assigned to English A this year, 1 among 158 was excused from the class with less than one term's instruction because that student seemed to have been unjustly held for the remedial course. This student was of foreign extraction and was compelled to cope with some language difficulty on account of the difference between a mother tongue and English.

The first half of the term's work in English A is devoted to a diagnosis of language difficulties. The instruction is intended to refresh old recollections of previous teaching and to increase the store of knowledge of technical English. Diagnostic testing is frequent and personal conferences are numerous. When a diagnosis is finally reached, the student is informed and is given work to do which is intended to remove his weaknesses. There seems to be a direct relation between the psychological test percentile and the learning power of the student, because the greatest improvement is usually observed among those students who rank highest on the psychological test, and the least improvement is usually observed among those students who rank lowest on the psychological test. This is a case in illustration: Miss K had a psychological percentile rating of 0.83, a high-school percentile rating of 0.44, and an English aptitude percentile rating of 0.15. After the diagnosis and remedial instruction, she rated equivalent to 0.96 on the English aptitude test. Or here is another typical case: Miss P had a psychological percentile rating of 0.26, a high-school percentile rating of 0.81, and an English aptitude percentile rating of 0.12. After the usual diagnosis and remedial treatment, she rated equivalent to 0.54 on the English aptitude test.

Additional data of similar nature are on file with the English bureau and may be consulted for further verification of certain parts of the conclusions about to be offered.

The following summary presents briefly the conclusions which we may reach at this point in our investigation of the prediction of success in college English composition. (*a*) The aptitude test now devised is thoroughly satisfactory for the purpose of segregating students for remedial instruction in technical English. (*b*) There is a direct relation between the psychological test score and the improvement from remedial teaching in technical English. (*c*) It would seem advisable to use the English aptitude test for the purpose of sectioning for regular college instruction those students not held for remedial instruction. (*d*) The use of the English aptitude test is in perfect harmony with a philosophy of education which sets forth the desirability of achieving maximum results at all levels of intelligence, whether that implies graduation from college or not.

REMEDIAL MEASURES FOR COLLEGE FRESHMEN

J. DeWitt Davis [1] and Harold Saxe Tuttle [2]

I. INTRODUCTION

The problem of college failure due to nonpassing grades has received considerable study, and several factors have been isolated (11, 13, 19, 20, 23, 28, 29, 30, 31, 37).[3] To remove these causes there are three general types of remedial work reported in the literature: (*a*) A careful study is made of each case, and work is assigned and supervised in such a way as seems best fitted to the individual involved. The program includes testing for diagnosis, conferences, some class work, and some individual coaching (27, 29). (*b*) Personal direction with no class treatment, depending largely upon the interview technique (17). (*c*) Group treatment, including more or less of the values apparent in the other two. This has generally been carried on through the medium of the so-called How-to-Study courses (5, 9, 15, 16, 25, 26). It is to this last group that the work at Oregon belongs.

II. EXPERIMENT IN REMEDIAL TREATMENT OF FRESHMEN

The work at Oregon might be termed preventive rather than remedial, for it is designed to anticipate the more common difficulties that have been found to exist in the work of beginning students and, by means of constructive reading, personal interviews, and student practice, to initiate such habits as may forestall maladjustment and eventual college failure. A 2-hour course is offered during the freshman year, under the title Freshman Orientation. This course was first offered in the school of education in 1927–28, and has been continued for two reasons: First, because of the values that it seemed to offer after the first analysis of results obtained; and, second, in order to accumulate further data that might be useful in directing the future course of such remedial work. The course is required of all freshman majors in education and during 1930–32 has been open as an elective to others.

[1] J. DeWitt Davis, Teaching Fellow in Education, University of Oregon. B. A., University of Idaho, 1913, M. S., 1929.

[2] Harold Saxe Tuttle, Associate Professor of Education, University of Oregon. B. S., College of the Pacific, 1905, M. A., 1911; B. D., Pacific School of Religion, 1911. He was formerly head of the department of education, Pacific University. Publications: With Harl R. Douglass, Project Teaching in a College Course in Educational Psychology, *Controlled Experimentation in the Study of Methods of College Teaching*. University of Oregon Publication, Education Series, 1:7:293-299, February, 1929; *Character Education by State and Church*, The Abingdon Press, 1930; with P. A. Menegat, *Procedures for Character Education*, Cooperative Store, University of Oregon, 1931.

[3] Numbers in parentheses refer to Bibliography, pp. 102-104.

The following outline will give a better idea of the nature of the work required by this course:

1. Synopses, readings, and class discussion on the wise use of time while in college.

2. The actual budgeting of time for various activities, with reports at frequent intervals of records of a week's time expenditure.

3. Habits of study. What are the physical requirements, external and internal conditions essential to good study? How can adequate study habits be built? What can one do to correct older habits that are not economical? How can one distribute his study over the day and week to secure maximum results?

4. Reading improvement. What causes poor, slow reading? Systematic records of regular drills for the measurement of improvement.

5. Planning how to increase interest in a given subject.

6. Three lectures on library procedure, each accompanied by a carefully assigned project, each project carefully marked for errors, returned, and required to be corrected by its author.[4]

7. Improvement of vocabulary. Value and methods. Readings, discussion, drills, and class quizzes.

8. Note taking, from reading and from lectures; readings and discussion with actual drill.

9. Suggestions for better reviewing; plans for review submitted and discussed, and later reports on how certain subjects were reviewed.

10. Study and discussion on how to keep fit physically and mentally, with help in self-analysis.

11. The importance of proper social adjustment.

12. Preparation for examinations; methods of preparation; types of examinations.

13. How to build up a good bibliography; actual drill required.

14. Preparation of term paper. Choice of subject, data, treatment, outline, mechanics of a good paper.

15. The physiology and psychology of learning. How are habits formed? Poor ones replaced by good ones. Aids to memory, proper distribution of drill, the value of appreciation, the place of imagination in adequate learning and living, reasoning in its various forms and applications, its common enemies. Exercises in self-expression, experimentation, observation of others, reading, taking notes, and making class reports.

16. The third term's work which is given to the education majors, but not included in that offered the nonmajors, consists in the main of a preview of the larger divisions of college courses from which

[4] Miss Casford, assistant librarian at Oregon, supervised this part of the program.

prospective teachers must select teaching norms. Reading drills and note-taking exercises are continued, their application being made to the following general topics: College requirements; physical sciences; biological sciences; social sciences; English; foreign language; physical education; music; philosophy; expressional activities, as public speaking, dramatics, story writing, and art. A provisional student program for the remaining three years of college work is planned by each member of the class.

This course must be carefully distinguished from survey courses in literature, natural sciences, and social science, which are commonly called orientation courses. In order to keep this distinction clear, Education 111, Orientation, may, for convenience, be called the How-to-Study course.

III. EXPERIMENTAL STUDIES

Table 1 indicates the enrollment during the 4-year period of the Oregon experiment.

TABLE 1.—*Enrollment in "how-to-study" course by years and psychological rating quartiles*

Class	0–24	25–49	50–74	75–100	Total
1	2	3	4	5	6
1927–28	6	6	8	9	29
1928–29	13	12	4	2	31
1929–30	17	7	9	7	40
1930–31	27	23	16	12	78
Total	63	48	37	30	178
Percentage in each quartile	35.4	26.9	20.8	16.9	100
In lower half	62.3 per cent		In upper half	37.7 per cent	

Each of these students has been matched, for the purpose of comparison, against another who is not taking the course. These controls are selected by the personnel department, care being taken not only to match them closely on the psychological entrance examination percentile score, but to pair them closely on the matter of high-school record as well, which record is also entered as a percentile score, making direct comparison easy. Because of some evidence that grading systems in general differentiate between men and women in favor of the latter (22), the pairing of cases has avoided this sex variability, matching only in the same sex. To illustrate how successfully this pairing has been done the following table has been prepared, which is a record of 76 cases who completed the first term of the course during the year 1930–31.

TABLE 2.—*Seventy-six paired cases of how-to-study students of 1930–31*

Quintile	Number	Experimentals		Controls	
		Psychological entrance examination	High-school grades	Psychological entrance examination	High-school grades
1	2	3	4	5	6
First	23	7.73	19.52	8.82	22.36
Second	21	29.09	38.05	29.14	38.76
Third	13	53.92	48.41	54.38	49.70
Fourth	11	69.29	43.60	68.91	42.20
Fifth	8	91.25	76.66	89.00	80.12
Average total	76	39.10	38.37	39.42	40.80

Particular emphasis is justified here, for this is the first extensive attempt to pair cases so exactly on both psychological scores and high-school grades. (High-school grades are regularly transmuted into a percentile preparatory rating by the personnel department.)[5] This care in matching cases has given direction to the study that would otherwise have been overlooked.

Considerable data have accumulated in this period of time, and they are being carefully analyzed in an effort to determine in how far there are statistically significant differences between the experimentals and the controls, and what the indications are that the how-to-study work was a causal factor in producing these differences; and further, to ascertain in so far as it can be done, how values derived from this course are manifested in later college work. Some of this analysis is now available.

On the basis of the first year's work, using average grades as a criterion, the results show as follows:

TABLE 3.—*1927–28 grade averages how-to-study and controls*

Term	Experimentals	Controls	Number
1	2	3	4
Fall	3.417	3.355	29
Winter	3.217	3.346	27
Spring	3.289	3.450	23

NOTE.—The data here reported vary in slight details from a later analysis of the same work due to the fact that here all the grades were used where later those of military, physical education, and personal hygiene were omitted.

[5] If psychological scores are interpreted as indicating native intellectual ability, high-school grades may be interpreted as reflecting, in considerable degree, habits of application, persistence, and effort.

The change of position from one of inferiority at the end of the first term, to one of considerable superiority by the end of the third looks favorable for the experimental group; but when this difference is treated statistically it loses much of its significance.

The change in relationship between the grades of paired cases is more significant. At the end of the first term there were 15 controls whose grade average exceeded that of their paired cases; by the end of the second term this number had dropped to 11, and at the end of the third to only 7. These results appear to indicate that the drill of the course, the suggestions for improved use of time, better reading habits, etc., gradually change the condition from one where the controls had a slight advantage to one where the experimentals had a clear margin in both grade average, and in a number of cases that exceeded their controls. This latter fact indicates the rather general effect of the work of the how-to-study group. These results appeared sufficiently positive to justify the continuation of the course. Data are therefore available for four years.

IV.—DATA AND ANALYSIS, 4-YEAR PERIOD

1. *Effect as shown by grade averages.*—Since the former workers have all utilized (5, 9, 15, 16, 17, 18, 21, 23, 25, 26, 29, 34, 40) the criterion of grade averages, in one form or another, that analysis was first made. The task is much more difficult and time consuming than the relating of it. To secure from the registrar's office, to compute, to tabulate, and to compare the work of 178 students, with from one to three similar control records for each of them, totaling 712 records in all, each of which records covers from 1 to 11 terms, is a very large undertaking. This work is not complete to date, but enough has been done to give some significant indications.

It is to be noted that grade averages have all been computed by omitting both military and those physical education and hygiene courses that are required of freshmen and sophomores. This was done arbitrarily, because the authors felt that grades in those courses would tend to obscure real differences that might develop otherwise. Since they are required of both groups, no injustice is done by omitting them in the analysis. Further, no weight has been given to grades marked incomplete, though a superficial analysis (Table 6) would indicate that were these "incompletes" included with their later assigned grade the average of the experimentals would be enhanced. The controls have somewhat more of such grades, and allowing a slight discount in grade value for tardiness, it is their average that would suffer. If an arbitrary value of say grade IV or V, or any other were used, the same condition would prevail for a like reason, hence they were omitted.

The data presented in Table 4 below were derived by grouping into one large distribution all the students' records for each respective term of college work that was completed to that date, counting as term 1 that school quarter in which the how-to-study course was first taken. The first three lines include the three years' records, 1927–28, 1928–29, and 1929–30, respectively. The fourth line is for 1930–31, and the fifth is a comparison of the last term's record of all the students who did the work in 1928–29 and 1929–30. The last line is a composite of all the grades involved in lines four and five.

TABLE 4.—*Differences in grade averages between how-to-study groups and controls*

[X = Experimentals; C = Controls]

Term	Number	Grade averages		Differences of mean C—X	D	Chance per 100 that true D is greater than zero[1]
		X mean	C mean		σ difference[1]	
1	2	3	4	5	6	7
1 1927–28	96	3.448	3.531	0.083	1.031	70
2 1928–29	89	3.301	3.428	.127	998	67
3 1929–30	76	3.315	3.562	.247	2.394	98
4 1930–31[2]	83	3.296	3.541	.245	2.764	99
5 1928–29 and 1929–30, last term	68	3.386	3.601	.215	2.370	98
6 Composite 1928–29, 1929–30, 1930–31	151	3.434	3.568	.134	3.098	100

[1] The formula used for this computation was the following:

$$\frac{D}{\sigma \text{ diff}} = \frac{D}{\sqrt{\sigma m_1^2 - \sigma m_2^2 - 2r_{12}\sigma m_1 \sigma m_2}}, \text{ where } \sigma m^2 = \frac{\sigma^2 dis}{N}$$

[1] Computed from Holzinger, Statistical Methods in Education, Table 42, p. 211.
[2] Same as footnote 1 in Table 5.

The numbers involved in each comparison are fairly large. In each grouping the experimental's mean grade average X (column 3), are consistently better than those of the controls C (column 4). This consistency of data tends to strengthen the probability that the cause of these differences in favor of the experimental group is more than mere chance, even though column 7 would allow some leeway for a chance factor in at least the first two terms. In the last line the latest available term's grades only were compared in each of the 151 paired cases. Of these some were first-term 1931 grades, some second, and so on as far as the eighth term. The difference in grade averages of columns X and C is small, only 0.134 grades, but the difference is clearly significant as indicated by column 7.

Another approach was made to determine whether the differences indicated were consistent from term to term for each year's students. Table 5 sets forth these average grade differences.

TABLE 5.—*Comparison of average grades—first three terms for 1927–28, 1928–29, 1929–30*

[X = Experimentals; C = Controls]

FIRST TERM

Year	Number	X	C	Difference
1	2	3	4	5
1927–28	27	3.575	3.400	−0.175
1928–29	30	3.492	3.708	+.216
1929–30	39	3.311	3.400	+.089

SECOND TERM

1927–28	25	3.385	3.335	−.050
1928–29	28	3.357	3.544	+.187
1929–30	36	3.271	3.403	+.132
1930–31 [1]	83	3.296	3.541	+.245

THIRD TERM

1927–28	21	3.387	3.411	+.024
1928–29	24	3.262	3.604	+.342
1929–30	31	3.310	3.480	+.170

[1] The last available grades of 1930–31 were compared; some were first term, some second term averages.

This table is consistent with that set forth above (Table 4). The negative difference in the first term's work of 1927 was gradually changed by the third term of that year into a positive advantage. The beginning negative difference may be due to the fact that the matching of pairs was not quite so thoroughly refined as for later groups.

To account for this consistent difference in favor of the experimental group the suggestion has been made that the content of the course, and the kind of treatment, is particularly valuable for prospective teachers, therefore they would naturally profit more than other students by it. Another suggestion has been offered, that a different type of student enrolls as an education major, and also the grading system of that department is different. Now, if such were the adequate explanation then theoretically some other group of non-education majors should not show these differences. This, however, is not the fact as indicated by items bearing footnote 1 in line 4 under second term of Table 5. Of this entire group, 83 in number, only 13 were education majors, the rest belonged to various other departments (except law). Yet the cases in this group show the largest difference of any group for the first or second term. Moreover, the difference as indicated by Table 4, is clearly significant statistically. This would seem to preclude all of the above objections and to point to the how-to-study remedial treatment as the causal factor. It is,

at least in this particular respect, that the groups are known to have been treated differently.

2. *Effect as indicated by differences in unsatisfactory grades.*—Since Jonas (17), Pressey (24), and Lemon (21) have all emphasized the fact that their experimentals showed fewer cases of unsatisfactory scholarship as indicated by failures, conditions, or probations, it was deemed worth while to analyze the data at hand for corroboration or negation of this emphasis. Table 6 sets forth facts discovered. The entire college record of 149 cases and their controls were available and were examined. Some of these records covered one term, others as much as 11 terms' work. In this analysis grades marked incomplete were included as unsatisfactory, on the theory that such a grade indicates some sort of maladjustment, or inability that the student was unable to remove on schedule time.

TABLE 6.—*Unsatisfactory grades of how-to-study students and controls*

Percentiles	Number of cases compared	Number of courses carried		Number of conditions incomplete and fail grades [1]		Per cent of such grades in each	
		X	C	X	C	X	C
1	2	3	4	5	6	7	8
0–24	50	629	628	44	61	6.99	9.71
25–49	39	380	370	21	16	5.52	4.32
50–74	33	518	503	20	11	3.86	2.12
75–100	27	561	575	7	13	1.24	2.26
Total	149	2,088	2,076	92	101	4.40	4.86

[1] These grades are not weighted for course hours. They are simply the number of such grades received.

NOTE.—Of a total of 4,164 courses only 193, or 4.63 per cent, were unsatisfactory.

In the whole group and in each of its quartiles the number of courses carried by the X and C groups are reasonably uniform, although the differences in the three upper quartiles may have had some bearing on the number of unsatisfactory grades (columns 3 and 4). The control group (C) had more unsatisfactory grades than the experimentals (X groups) in the first and fourth quartiles, and in the total. Though this difference is not as large as Book (5) reported, it does indicate a similar condition.

These data also emphasize the importance of the statement that there are other factors causing dropping out of college, for of a total of 4,164 courses taken only 4.63 per cent were not completed satisfactorily. Lemon (21) presented data that showed 57 per cent of the lowest decile as dropped out by the end of the first year. Here it is indicated (columns 7 and 8) only 7 to 10 per cent of the 2,257 grades turned in for the entire low-quartile group are of unsatisfactory quality, and some of these records reach as far as the fourth year's work.

Closer study of this table (columns 5 and 6) suggest that students respond to the type of work offered in the how-to-study treatment in a different way, depending somewhat upon their ability and habits as indicated by their psychological and preparatory scores.

Both the lowest and the highest quartile show marked favorable difference; the low-group controls receiving 34 per cent more unsatisfactory grades, and the high-group controls getting 85 per cent more such unsatisfactory grades than the corresponding groups of experimentals. Yet the data is not consistent, for the middle groups favor the controls in this respect.

Another approach gave a more promising lead, and one that throws new light on Pressey's recent statement to the effect that students above the twenty-fifth centile in ability profit most from such remedial treatment. In the preliminary analysis of 76 subjects carefully paired with controls there was some evidence that differences between percentile ranking in preparatory grades and psychological scores were more significant with respect to improvement under guidance than were either of the scores taken separately or the two combined.

This would appear to mean that students whose achievements in high school were distinctly lower than their psychological tests would lead one to expect were helped quite decidedly. Students whose achievements in high school were higher than their psychological score would lead one to expect were aided but little by the how-to-study course.

This clue led to a fuller study of the relation between the differences of percentile ranking and scholastic attainment.

A study of the effect of percentile differences brought to light the following facts: When the average grades of all students in the lowest quartile were compared with those of their controls, disregarding the matter of spread between the two percentile scores, it was found that there was only a slight difference—namely, 0.011 grade value—in favor of the experimental group. This difference has no significance statistically and appears to corroborate Pressey's conclusion (26) with respect to low-quartile students, suggesting that work with them is too expensive to be justified in the light of results secured.

However, 55 other case records with their controls were studied where the psychological score was 20 or more centiles below the preparatory score (which, as already stated, may be thought of as a habits-of-study score). The mean was 0.077 grades in favor of the experimental group. This difference also is not significant statistically, though it is consistently favorable to the experimental group. However, if only ability, as indicated by the psychological scores is important, then this difference should be sufficiently greater than that noted above to be significant, for the group includes a large number whose rank is well above the lowest quartile. It appears that those students whose habits of study, as evidenced by their

preparatory score, rank above their own ability level, as indicated by the psychological score, are least affected by the how-to-study remedial work.

This surmise is further strengthened by the data presented in Table 7, wherein the records of 151 cases and their controls were carefully analyzed. The last available term's averages were used. These cases were divided into two large groups included in the data presented in lines 4, 5, and 6 of Table 4. In each case the differences for the separate groups were highly significant and when massed into one composite group of 151 cases have a surplus of significance.

Each of these 151 cases was put into one of the following categories:

1. Those whose psychological score is 6 centiles or more greater than their preparatory score.

2. Those whose psychological score is within 5 centiles of their preparatory score.

3. Those whose psychological score is 6 centiles or more less than their preparatory scores.

The basic assumption involved here is that the psychological score is an ability index, and that the preparatory score is an index to habits of satisfactory adjustment; which habits it is the purpose of remedial work to build up. Theoretically, then, those students who are already working above their ability level should be helped the least.

As one examines the data set forth in Table 7 it becomes apparent that, in so far as the experiment has progressed, the early surmise is justified that those students whose achievement record evidenced by preparatory scores is lower than their ability are aided most by the remedial work; and that those students who are already working over their ability by the same index when they enter the course are helped least—if, indeed, at all.

TABLE 7.—*Comparison of how-to-study and controls in relation to psychological and preparatory scores*

Year	Group	Psychological score greater than preparatory		Psychological score equal to preparatory		Psychological score less than preparatory	
		Number	Average grade	Number	Average grade	Number	Average grade
1	2	3	4	5	6	7	8
1928–29	X	12	3.312	17	3.389	39	3.400
1929–30	C	12	3.500	17	3.889	39	3.516
Difference			+.188		+.500		+.116
1930–31	X	28	3.616	26	3.605	29	3.676
	C	28	3.812	26	3.472	29	3.533
Difference			+.196		+.123		+.143
Composite of both groups	X	40	3.512	43	3.520	68	3.518
	C	40	3.719	43	3.668	68	3.459
Total difference			+.207		+.148		−.059

This modifies the conclusion of Pressey (25) that the ineffectiveness of remedial measures is due specifically to low intelligence, implying rather that it may be due to the fact that all with low intelligence who gain admission to college have already developed habits of study superior to the average of their abilities.

V. READING IMPROVEMENT

From the beginning of each year's work, considerable emphasis was placed upon the importance of efficient reading. What are the mechanics involved in such reading? What are the chief enemies to good reading habits, and how may they be overcome? The problem here implied was made a matter of major study and drill.

Careful records for three groups have been made from week to week throughout the period of drill. Each student, three times a week, reads some uniform material, at least 20 minutes, making an effort to apply the ideas gathered about efficient rapid reading, not skimming. From each of these efforts he makes a words-per-minute reading check. Practically complete records for one term were secured from 70 students. The master group sheet shows only the weekly average of these three or more records.

To stimulate interest in reading improvement very specific action was followed. The class was given reports showing actual improvements made by similar groups. This was done every week at the second session. The records were received the first session. The exact average of the group, the median, the high, and the low, were also given on the blackboard in table form and each student was urged to keep his own record, indicating where in this total group he found himself from week to week. Moreover, from time to time short formal and informal reading tests and speed checks were given in class period. This served the double purpose of added group drill and of added records to compare with student reports.

On the whole the data show positive gains not identical with, but comparable to, and in general corroborative of other remedial work in improving reading, such as that of Book (4), Remmers (30), Pressey (27), and others. Table 8 sets forth some of these data, which were gathered from three separate groups, for convenience called A, B, and C. A was a class of 33 noneducation major freshmen doing the work here reported in the 1930 fall term. B was a group of winter term noneducation freshmen 14 out of 20 belonging in the low half of ability, including 5 belonging in the low decile. C was a group of chemistry students. Because of schedule difficulty this class (C) was divided, allowing more than the usual amount of personal attention to individual difficulties. When the data of these three classes are thrown into one distribution, it tends to show a more reliable picture. It is clear, however, from Table 8 that Group B was benefitted least

by the reading drill. The fact that Group B was not given individual interviews involving reading diagnosis and suggested changes for improvement, as were the other groups, may account for some of the failure to respond as well as did both other classes.

TABLE 8.—*Reading rate improvement of three how-to-study groups, 10 weeks*

Group	Number	Quartile	Average words per minute			Quartile per cent gain	Group gain
			First week	Last week	Gain		
1	2	3	4	5	6	7	8
A	13	0–24	239	297	+58	+23.9	
	9	25–49	278	371	93	33.2	[1] 57
	7	50–74	262	343	81	30.9	[2] 26.27
	4	75–100	309	343	34	11.2	
B	6	0–24	218	232	+14	+6.4	
	8	25–49	232	276	44	19.0	[1] 45
	5	50–75	176	233	57	32.3	[2] 21.43
	1	76–100	158	280	122	77.21	
C	8	0–24	230	373	+143	+61.7	
	4	25–49	222	335	113	51.3	[1] 109
	1	50–74	173	262	89	51.4	[2] 44.46
	4	75–100	319	364	45	13.8	

[1] Words per minute.
[2] Per cent.

Table 9 shows the gains in reading speed made by each quartile when the data was massed. Emphasis was placed on improved rate, it being assumed from former studies (1, 30, 33) that comprehension follows closely with the increased rate.

TABLE 9.—*Reading-rate improvement—10 weeks record of 70 cases*

Percentile	Number	Average words per minute			Per cent gain
		First week	Last week	Gain	
1	2	3	4	5	6
0–24	27	232	305	73	31.5
25–49	21	250	324	73	29.3
50–74	13	206	273	67	32.6
75–100	9	296	344	48	16.4
Total	70	240	312	72	30.0

Some facts stand out clearly in these figures. First, as a group it read about 6 words per minute too slowly for college freshmen (38) when first measured. Second, as a group under the remedial treatment given, it responded with a 30 per cent increase in speed arriving at 66 words per minute advantage over the norm of 246 words as established at Nebraska by Werner in 1926. Third, lack of personal interview or some other cause or combination of causes resulted in

less improvement on the part of Group B. Fourth, all those students who appeared to be in earnest and interested in doing so succeeded in increasing their reading speed, while others with plenty of ability such as case Go, rating 66–62, began at 203 words per minute and ended at practically the same point, 214. Another case Ek, rating 91–91 in psychological and preparatory scores, respectively, began at the very inefficient level of 158 and ended at 280 words per minute; and Ok, rating 26–02, began at 173 and ended at 340, showing a gain of 167 words per minute, or 98.2 per cent.

Figure 1 has been prepared to indicate the spreading effect of remedial work in reading. It is of interest to note that the upper limit graph is the score made by the same individual while the honors at the bottom of the list, as would be expected, were shared by several different students from week to week.

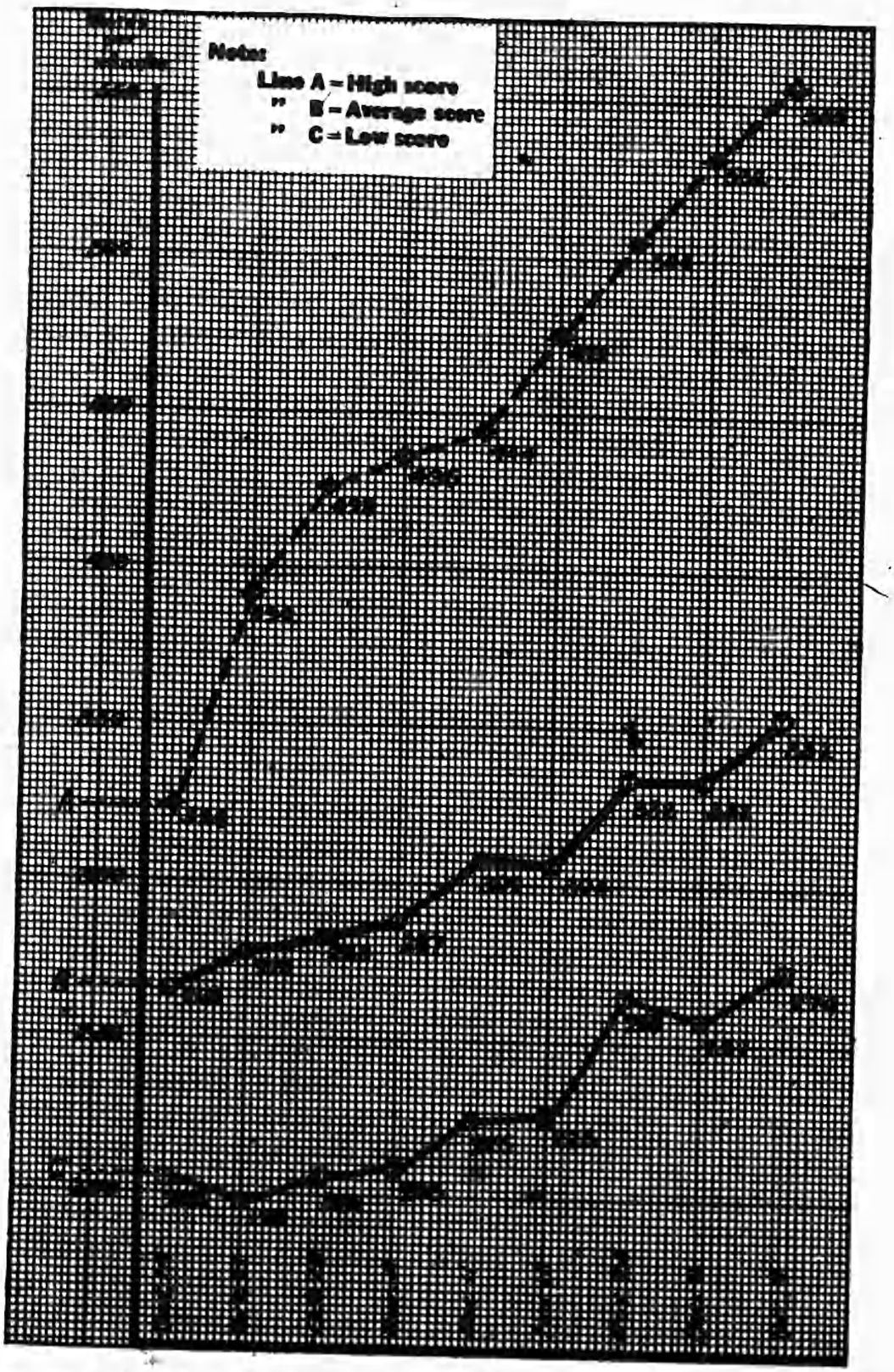
While there was considerable difference between the high and the low reading rates the first week (119 words per minute), yet the deviation from the mean was less than that of later records. The slow readers appear to be working hard to keep up, and the more capable ones appear to have been loafing. The slowest reader was 63 below the mean and the best one only 56 above.

When reading drill began this spread at once started to grow. The average of the whole class increased steadily, but the greatest speed records were made by those who had the higher psychological ratings. (See Table 9.) At the end of the 10 weeks' period the spread between the high and low per minute records had increased from 119 to 281 words, showing a 236 per cent increase in spread, apparently due to the training. Of this total spread the greatest deviations were regularly above the mean of the group; so that in the last record, the distance from the average rate of the total group, to the lowest score is now 78, as compared to 68 words per minute for the first week's record; and the distance above the average to the highest score is increased to 203 words per minute, where it began as only 56.

Massing data as in Table 9, or graphing it as in Figure 1, reveals larger tendencies which are indicative of value to be derived from reading training, but it tends to obscure more detailed facts such as those set forth in Table 8.

Column 5 of Table 9 indicates that ability quartile one, ranks second in the total group for words per minute gain in reading speed. The numbers involved are too few for any extended analysis, but it is interesting to observe in Table 8 that quartile one in Group B showed an average gain of only 14 words, in Group C the best gain of any, 143 words, and in Group A, a gain of 58 words, only 14 lower than the average gain for the whole group. In the light of this extreme variation, when one recalls the difference in treatment given each group as before noted, is there not an implication that improvement in read-

ing rate depends not entirely upon ability as indicated by the psychological scores but rather both upon the amount of efficiency when



the work began and upon the kind of remedial treatment offered? This implication would place new responsibility upon the colleges for

both more adequate diagnosis and teaching, and allow little ground for blaming lack of improvement in reading entirely to lack of ability. In short, the top and bottom quartiles appear to need a different type of remedial treatment which, when given to each in the best way, may perhaps tend to cause each to approach its maximum capacity in reading rate.

To test this hypothesis the whole problem of reading improvement should be restudied, first on the basis of differences between ability (psychological) scores and habit (preparatory) scores, and second on the basis of differentiated treatment for the lowest and the highest ability groups.

It would seem that to utilize this ability made available in such remedial work, the length and difficulty of assignments involving considerable reading should be regulated so as to provide not only for the poor readers where it appears frequently to be placed, but also to demand from the better students more of their ability made available by such training as here provided.

The question is asked, will this increase become a permanent possession? This question can be answered theoretically by putting another question. Reading is a skill habit; will any skill habit, e.g., typewriting, shorthand, or piano playing, be retained unimpaired if it is not used regularly? Will it be used regularly if assigned tasks do not demand it? Does one exert himself and maintain high efficiency when necessity does not require it? Does college study put the most rapid readers on their mettle?

VI. VOCABULARY DRILL

Words are tools to aid in shaping social adjustments. Words become surrogates for large blocks of past experience when one learns to use them efficiently. In this sense words are keys to the treasures of the past, and talismans to the secrets of the future. Upon such a theory of word value vocabulary drill has been recently included in the how-to-study remedial program.

Lack of word understanding manifested itself in different ways. In the personal interviews related to reading diagnosis several subjects, when asked why their eye movement regressed, replied that they did not get the meaning involved because of some new word or word usage.

It has been well established that regressive eye movements are one of the several causes of inefficient reading. One may conclude then that as a part of the reading improvement such vocabulary drill should be encouraged, for as words become familiar proper meaning is derived more quickly and less flitting back of the eye is necessary.

Somewhat of this lack of word understanding may be indicated by quoting a few student responses given on various quizzes:

Choleric: Being ignorant about.
Chimerical: Quick tempered.
Cozening: To reason earnestly.
Gregariousness: Quarrelsome.
Ambiguous: Gigantic.
Fallacious: Condescending.
Ephemeral: Effeminate.
Decorum: Belief.
Expostulate: To eradicate.
Hedonic: Unfortunate.
Grovel: A mere trifle.

In fact, a collection of some of the student choices in best meaning put together would be fair material for "College Humor" or "Facetious Fragments."

Lists of 70 words were assigned each week. Students were advised that a random selection of 20 words from these lists would be given in a quiz every 2 weeks. The type of quiz was explained, and a sample given, all of which placed the emphasis not upon single synonyms, but upon larger meaning content and usage, for the student never knew just what part of the meaning pattern would be used in the quiz.

An effort is being made to determine the relation of the scores made on the vocabulary tests to the average term grades of each student and to the grades in specific subjects as compared with the control averages and grades. This analysis is not complete, however, at this time. As indicated by the appraisal of value by students it would seem to be a useful form of remedial treatment.

VII. STUDENTS' APPRAISAL OF VALUES

In Ohmann's study (23) the suggestion is offered that the effectiveness of the treatment given throughout his course could not be accurately measured until some time later, but he added, "A subjective conviction of its value came perhaps most forcefully from the expressed appreciation of individual students who had been helped." In this Oregon study definite effort was put forth to secure and tabulate such student expression and to evaluate it.

Procedure: At the close of the winter term the following blank was put in the hands of each student who had been in this course since it was first offered in 1927.

COPY OF LETTER FORM

(School of Education letterhead)

March 3, 1931

No. ------

One time member of Orientation Class, Education, 111.

DEAR FRIEND:

We are making a survey of the opinion of various students as to what each one thinks are the most lasting values derived from certain courses in college work. You have done some study in a Freshman Education Course No. 111, called Orientation. In view of its bearing, as you look over your whole college career, can you name five elements, things, or phases of that course from which you derived some value which was of more or less help either in your studies or in your general adjustment to daily life? Now as you look back over that course,

select the one of the five which you rate at present of greatest worth, and assign it the value 5; similarly score the others in descending order, assigning the value 1 to that which you estimate as lowest.

Your reply will be treated as confidential; your name need not be signed unless you prefer to sign it.[1] Prompt return of this sheet will be appreciated, and will expedite our study.

Write your five items here:

Assigned value	

Further remarks: ----------

Sincerely,

The form was carefully prepared to avoid any suggestion, and yet to require emphasis only on the values which were outstanding to the student at the time. Only five such values were asked for, thinking that such a limitation might make the emphasis more indicative of actual conditions.

Returns were received from 97 students, some having had the work of the course as early as 1927–28. Data were tabulated as indicated in Table 10.

TABLE 10.—*Student evaluations of the course*

Values indicated by students	Relative rank	5 first place	4 second place	3 third place	2 fourth place	1 fifth place	Per cent of replies	Weighted values *
1	2	3	4	5	6	7	8	9
Training and drill in scheduling time for study, analysis of time expenditure	1	24	23	11	9	5	74	268
Improving reading ability	2	17	22	20	9	8	78	259
How to use the library more efficiently	3	10	12	4	10	10	47	140
Interest in increasing one's vocabulary	4	9	5	10	12	18	56	137
Study improvement learned how to improve my methods of study	5	12	6	4	1	3	28	103
General content of reading material assigned	6	10	3	3	4	3	24	80
How to make better notes and how to use them	7	3	3	5	2	3	20	55

* The weight values in column 9 were computed by finding the sum of all of the products of the number of students by the weight assigned to each value indicated.

[1] Most all of the returns were signed.

TABLE 10.—*Student evaluations of the course*—Continued

Values indicated by students	Relative rank	5 first place	4 second place	3 third place	2 fourth place	1 fifth place	Per cent of replies	Weighted values
1	2	3	4	5	6	7	8	9
How to review and prepare for quizzes and examinations	8	2	3	8	4	1	19	55
Something learned about the physiology and psychology of college work	9	3	1	5	4	6	20	48
Unclassified miscellaneous values	10	2	4	4	3	5	19	45
Drill in mathematical problems [1]	11	3	2	3	2	1	11	37
Survey of different fields of possible college courses to follow	12	0	1	4	3	3	11	25
Was of general aid in all college work	13	0	2	3	2	2	9	23
Drill work on equations and formulas [1]	14	1	2	1	2	1	7	21
Guided toward vocational choice, or aided in knowing how to choose	15	1	0	3	2	0	6	18
Preparation of term papers, technique, etc	16	0	2	2	0	2	6	16
Personality improvement	17	1	1	1	0	1	4	13
Use of memory aids, etc	18	0	1	2	1	0	4	11
Created interest; learned how to increase interest	19	0	0	3	0	0	3	9
Taking part in discussion	20	1	0	0	1	1	3	8
Cultivated habit of promptness	21	0	1	1	0	0	2	7
Improved my mental health	22	0	0	1	0	1	2	4
Learned how to compile a bibliography; the mechanics of it	23	0	0	0	1	1	2	3

[1] These replies all came from the chemistry how-to-study section.

Value 1.—The item receiving first place, mentioned by 74 per cent of the returns, relates to economies in the use of time. The value, as they indicated it, being in actually scheduling one's own time, rather then merely in reading about how it is done, or in discussing the subject at class period, e. g., subject 19 gave "Planning for the week's work" second place and put "improvement" in parenthesis, indicating that he had noted his own progress. No. 9 of the 1929–30 class gave "arranging a study schedule and having definite study habits" first place. Another student (No. 36 of the 1929–30 group) said: "Taught me the value of a definite schedule, not only in school but also out."

If this item is of chief value, as seems here indicated, this fact may in part account for the similar results that are achieved under apparently different treatment. Essentially every attempt at remedial work has emphasized this phase of its program. L. Jones (17) gave help on the basis of "constructive individual guidance—without waiting for difficulties to arise to initiate such assistance." A major part of this guidance is related to economies in the use of one's time, as is indicated in the following paragraph:

> The time charts, used to reveal to the student and to the writer the amount of time actually given to studying, enabled the writer to advise discriminatingly, and the student to schedule adequately, the amount of time needed for his studies in order to improve his record. Of the four variables—native ability, time, study methods, and grades achieved—in a student's career, he can exercise more control over the use of his time than over any other one. * * * (17).

As measured by the five criteria selected, Jones found his experimental group significantly superior in each respect to the controls. His emphasis is the same as that at Oregon with respect to this value; the treatment is different, his being the more expensive case method, the method here that of group direction. Though the methods differ, there is a common point of emphasis on self-analysis in the expenditure of time which seems to contribute toward similar results in improvement, and which impresses the students with its value as herein indicated.

Value 2.—In the light of the experiments that have been reported relating to reading improvement, the ranking of this item in second place is easily understood. More students included it as one of the five chief values than any other item; but more of them placed it lower down in the scale, second or third, so that when properly weighted it became a close second—259 as compared to 263 for value 1. The improvement in reading, as set forth above, offers some explanation of this student evaluation.

Value 3.—As estimated by these student reports, the instruction and drill in how to use the library more efficiently receives third place in value. This is not to be wondered at when one notes the tendency in college courses to depend more and more upon current magazine material, and a diversity of authority rather than upon some specific textbook (10, 14).

Value 4.—The high rank of vocabulary training as a value derived from the course may require elucidation. In the first place the 1930–31 chemistry group [6] reported with strong emphasis here, 11 out of 16 valuing this part of their work highly. In the second place, during this year greater emphasis in all of the how-to-study classes has been put upon actually building up one's vocabulary, while formerly only what might be done as suggested by various authors, was pointed out. At three times during the work the students were called upon to hand in lists of words which they had found in their reading and had added to their vocabulary. From these lists quizzes were prepared which stimulated the student to further effort. This greater emphasis may account for the fact that vocabulary building was assigned fourth place in the total group of responses, whereas it was not accorded any importance by the replies coming from classes prior to the last year when drill on vocabulary was introduced.

Value 5.—By some other more or less arbitrary grouping of replies, putting every one that related in any way to improvement in study habits, learning how to study, better study as related to notes, lectures, reviews, general health, etc., under this heading, one might

[6] In the 1931 winter term an effort was made to give a group of chemistry students the same sort of remedial treatment, but focusing the drill specifically on chemistry content, e. g., library work was assigned on topics of interest to chemistry students, reading was encouraged on chemistry material, chemistry problems were assigned for fundamental arithmetic drill and vocabulary building was all in the field of chemistry.

have found this item in the most important position. By classifying many of such more general replies under other appropriate headings the emphasis still holds this item to be one of the first five values derived from the course.

Miscellaneous; unclassified values.—It has been suggested[7] that there may be a delayed benefit possible in such work. Some of the replies classified here as miscellaneous add weight to this suggestion. No. 45, now a substitute teacher, places the following emphasis: Value 5. The course "enables a grade teacher to begin (pupil) study habits correctly." Value 4. It "gives the teacher better understanding of a pupil's errors." No. 73, now a teacher in an elementary school, reports, "I find that the study I did * * * has been a great aid to me. I am in favor of the * * * courses, as I think they help one a lot."

Another student, No. 55, attributed greatest value to personal conferences and advice received in them.

The remarks written in on the returns from the students of former classes were, save for two exceptions, strongly in favor of offering such a course to all freshmen. Some of them are given below.

TABLE 11.—*Typical remarks from how-to-study students*

Case No.	Ranking		Remarks
	Psychological	Preparatory	
7	98	85	"I think the course was very beneficial to me and I think it would be good for most freshmen."
26	72	61	"I have been able to plan ahead. * * * Do not leave it out."
2	67	66	"It is a course which I believe would be beneficial to every freshman."
37	23	39	"Every college freshman should take the course."
6	17	96	"Orientation is a help to any student as we are taught how to save our time and study most effectively."

The psychological and preparatory percentile ranks are given to suggest that these comments come from various levels of student ability and training although in general the tendency as here evidenced, is for the upper half to recommend such work oftener than the lower half do. This may possibly mean that it is they who derive the most lasting benefit from the course as it is now being taught.

VIII. SUMMARY AND SUGGESTIONS

1. *Summary.*—A review of the reported work designed primarily to enhance the chances of college success for freshmen shows consistently positive results, as measured by various criteria.

The technique of remedial work varies. One type is essentially that of personal tutoring. To this method some would raise the

[7] Edward S. Jones, director of personnel research at the University of Buffalo, Buffalo, N. Y. In a personal correspondence, dated Jan. 24, 1930, to Dean H. D. Sheldon at Oregon.

question as to whether the results achieved are worth the expenditure of time and money involved. Another type is that of personal guidance illustrated by the work of L. Jones at Iowa. A third type which has been employed most extensively, and to which the work at Oregon belongs, is that of group treatment. This type generally includes in its program personal interviews and tests both for diagnostic and motivation value, as do also the other two types of program. Reading, discussion, some lectures—generally combined with drill in note taking—and a large amount of specific drill are also typical of this last group.

Thus in content the treatment given by the three is largely the same. Jones at Buffalo differs in his group work by crowding the remedial treatment into a preregistration period. This procedure would seem to have some advantages and some drawbacks. It would allow all attention to be centered upon one thing, namely, remedial work. It would also be easy to segregate the expense and charge it to the students who do the work. Being of a nature which perhaps should have been mastered before college matriculation this may be a justifiable procedure. It would also, as Jones points out, tend to allow a few of the very least capable to drop out before registration.

No reports of results achieved from purely lecture courses on how to study, were found in the literature. The authors of this report are utterly skeptical about such courses, if there be any; for habits are built or corrected not by exhortation or by telling why and how, but by actually doing, by personally experiencing good methods of study, by drill in better ways of doing college work, and by measuring one's own progress in this development.

In this report of the how-to-study work, members of experimental groups are shown to have a consistent superiority over the control groups, when this superiority is measured by average term grades. It seems from the data analyzed that this advantage is due at least in part to the treatment received in the how-to-study program.

Both the analysis of average grades from term to term and of the students' own statements as summarized in Table 10 indicate that some values are retained for later use. Table 4 indicates that average grade superiority manifested itself more significantly the third term than the first or second. But the most significant advantage was apparent when the average grades for the last available term's record, some as late as the first quarter of junior work, were compared. The how-to-study group also made slightly fewer grades of condition, incomplete, or failure than their controls and continued to do so from term to term, though these grades are too few in number to account for any large part of college mortality.

Further analysis of the data as indicated in Table 7, column 8, shows that it is not merely the low quartile ability group that profit

little or nothing by the how-to-study treatment; it is rather those who appear to be working above their ability when they enter the course. This finding may prove to be of value to personal advisers for interviews and guidance.

The results of reading drill are in general corroborative of those obtained in other studies; but the facts set forth indicate that to obtain a reading improvement at the different levels of ability the teaching and general drill method must vary, the low group requiring specific, detailed patient drill, where mere suggestion may suffice for the best students. The spreading effect of this reading drill is conspicuous. In the work two types of motivation appear of especial value; first, personal difficulty diagnosis, and, second, regular notation of the improvement of the group as a whole, indicating the best and the poorest score from week to week.

No analysis as to the effect of vocabulary drill is complete enough to indicate measurable values. Student emphasis ranks this work as being worth while.

Two points stand out clearly with respect to drill in library work: First, that two departments can cooperate effectively in such a program of remedial treatment; second, in personal interviews and in the replies tabulated, library drill was given a high rank as having value for student success.

The student replies may also be of service in suggesting where emphasis may be placed in remedial work with promise of greatest returns.

One major conclusion is evident from the data presented, namely, whether determined by objective statistical treatment or by subjective student evaluation, the parts of the remedial how-to-study program that seem to contribute most toward student success, consist largely of those things which the student actually does, drills at, experiences, and notes progress in. At this point all studies agree thoroughly.

In general it appears that the how-to-study work should be continued at the college level and made available to all freshmen, but that those freshmen particularly whose psychological scores are within a few centiles of or greater than their preparatory scores should be encouraged to elect the work of this course.

2. *Suggestions.*—Out of this study have arisen some problems which seem to merit some further experimentation.

The remedial work for chemistry majors along similar lines to those followed in the 1931 winter term might well be continued far enough to establish its value or lack of it. This could be done by advising its election in the fall term on the part of freshmen whose psychological scores are distinctly higher than their preparatory scores.

Considerable evidence indicates the value of the personal interview for both diagnostic and remedial work. To secure maximal values personal interviews require freedom from interruption and considerable privacy. Provision for both of these factors would no doubt increase the value of the work.

A study of the relation that may exist between hours spent in study per week, reading rate, vocabulary test scores, and psychological and preparatory scores is now being made by the authors. It may shed some further light on the problem of diagnosis and remedial treatment.

Vocabulary training with a large group in some specific field like chemistry or biology following the technique used in this study, with cases as carefully matched, is another promising lead for further research.

The implication of this type of remedial work is an old one—that college education should be not faculty centered or curriculum centered but student centered, with a program in which all are encouraged to function at their best.

IX. BIBLIOGRAPHY

(1) Alderman, Grover H. Improving Comprehension in Ability in Silent Reading. Journal of Educational Research, 13:11–21; January, 1926.

(2) Arnold, H. J. The Standing of College Students in Two Elementary School Subjects. *In* Research Adventures in University Teaching, by S. L. Pressey and others, pp. 107–112. Public School Publishing Co., Bloomington, Ill., 1927.

(3) Averill, L. A., *and* Mueller, A. D. The Effect of Practice on the Improvement of Silent Reading in Adults. Journal of Educational Research, 17:125–129; 1928.

(4) Book, Wm. F. How Well College Students can Read. School and Society, 24:242–248; August, 1927.

(5) ——— Results Obtained in a Special "How-to-Study" Course Given to College Students. School and Society, 24:529–534; Oct. 22, 1927.

(6) Cole, L. W. A Partial Remedy for Loafing in College. School and Society, 20:311–313; Sept. 6, 1924.

(7) Colvin, S. S. The Use of Intelligence Tests. Educational Review, 62:134–148; September, 1921.

(8) ——— Educational Advice and Direction of College Students. Christian Education, 5:18–35; October, 1921; July, 1922.

(9) Crawford, C. C. Some Results of Teaching College Students How to Study. School and Society, 23:271–272; Aug. 10, 1926.

(10) English, A. J. How Shall We Instruct the College Freshman in the Use of Library? School and Society, 24:779–785; Aug. 6, 1927.

(11) Germane, C. E., *and* Germane, E. G. Silent Reading. Row, Peterson & Co., 1922; p. 1.

(12) Guiler, W. S. A Program of Diagnostic and Remedial Instruction. The American Association of Teachers Colleges Yearbook, 1927; pp. 39–50.

(13) Harrington, M. S. The Problem of Mental Hygiene Courses, for the College Student. Mental Hygiene, 11:536–541; 1927.

(14) Humphrey, Gambier-Bowsfield. Do You Know Your Library? Journal of Educational Sociology, 4:93–104; October, 1930.

(15) JONES, EDWARD S. Testing and Training the Inferior Freshman. Journal of Personnel Research, 5:43–ff; 1926–27.

(16) ——— Preliminary Course in How to Study for Freshmen Entering College. School and Society, 29:702–705; Jan. 1, 1929.

(17) JONES, LONZO. Personal Service and Freshman Scholarship. Educational Record, 12:1; January, 1931.

(18) KITSON, H. D. The Scientific Study of the College Student. Psychological Monograph, 23:1–89; 1917.

(19) LAIRD, DONALD A. A Study of Some Factors Causing a Disparity Between Intelligence and Scholarship in College Students. School and Society, 19:290–292; Mar. 8, 1924.

(20) LEATHERMAN, ZOE E., *and* DOLL, E. A. A Study of the Maladjusted College Student. Ohio State University Studies, 2:2; 1925.

(21) LEMON, ALLEN CLARK. An Experimental Study of Guidance and Placement of Freshmen in the Lowest Decile of the Iowa Qualifying Examinations. University of Iowa Studies in Education, 3:8; 1925.

(22) LENTZ, T. F. Sex differences in school marks with achievement test scores constant. School and Society, 29:65–68; Jan. 12, 1929.

(23) OHMANN, OLIVER ARTHUR. A Study of the causes of Scholastic Deficiencies in Engineering by the Individual Case Method. University of Iowa Studies in Education, 3:7; Jan. 15, 1927.

(24) PAYNE, W. L. Methods in Teaching How to Study. School Review, 38:598–604; October, 1930.

(25) PRESSEY, L. C. The Permanent Effects of Training in Methods of Study on College Success. School and Society, 28:403–f; Sept. 29, 1928.

(26) ——— A Class of Probation Students at Ohio. *In* Research Adventures in University Teaching. Public School Publishing Co., Bloomington, Ill., 1927; pp. 134–139.

(27) ——— *and* S. L. Training College Freshmen to Read. Journal of Educational Research, 21:3; March, 1930.

(28) ——— Background Educational Factors Conditioning College Success. Studies in Educational Yearbook No. 16 of The National Society of College Teachers of Education, pp. 24–29. The University of Ohio Press, 1928.

(29) REMMERS, H. H. A Diagnostic and Remedial Study of Potentially and Actually Failing Students at Purdue University. Bulletin of Purdue University No. 28, May, 1928, Studies in Higher Education 9. The University of Iowa Press, Iowa City, Iowa, 1928.

(30) ——— *and* STALNAKER, J. M. An Experiment in Remedial Reading Exercises at the College Level. School and Society, 28:730, 797–800; Dec. 22, 1928.

(31) RUGGLES, ARTHUR H. College Mental-Hygiene Problems. Mental Hygiene, 9:261–672; April, 1925.

(32) SEATON, J. T. The Errors of College Students in the Mechanics of English Composition. *In* Research Adventures in University Teaching, by S. L. Pressey and others, Bloomington, Ill. Public School Publishing Co. 1927; pp. 96–99.

(33) STONE, C. W. Improving the Reading Ability of College Students. Journal of Educational Method, 2:8–23; September, 1922.

(34) ——— How to Study as a Source of Motive in Educational Psychology. Journal of Educational Psychology, 11:348–354; 1920.

(35) STRONG, RUTH. Another Attempt to Teach How to Study. School and Society, 28:461–f; Oct. 13, 1928.

(36) TOUTON, FRANK C. Report on Certain Phases of the Educational Guidance Program now in use at the University of Southern California. Phi Delta Kappa, 8:24–34; July, 1926.

(37) UHRBROCK, RICHARD STEPHEN. The Freshman's Use of Time. Journal of Higher Education, 2:137; March, 1931.

(38) WERNER, O. H. Every College Student's Problems. New York, Silver Burdett, 1928; p. 166.

(39) WHITE, C. L. The Freshman. Educational Administration and Supervision, 12:95–104, February, 1926.

(40) WITTY, PAUL A., *and* LEHMAN, HARVEY C. Teaching the College Student "How to Study." Education, 48:47–56, September, 1927.

(41) WOODRING, WM., *and* FLEMMING, C. W. Diagnosis as a Basis for the Direction of Study. Teachers College Record 30:46–64 and 30:124–147; October, 1928; November, 1928.

GROUP III.—ADMINISTRATIVE MEASURES BASED UPON TEST RESULTS

AN APTITUDE TEST AS AN AID IN ADMINISTERING LARGE SECTIONED COURSES

A. B. STILLMAN [1]

INTRODUCTION

For several years those in charge of the course in constructive accounting at the University of Oregon have been making a conscious effort to study out proper methods of instruction and accurate measures by which to gage the student's progress.

During the course of this program of experimenting considerable time was given to the working out of an aptitude test which would predict within the field of accounting more accurately than the general psychological test given to all students entering the university. Such a test was finally devised. While it has never been used as a sole basis for judgment of a student's aptitude or for assigning him to a particular section, it has proved of a great deal of value in a number of ways as an aid in the administration of the course. The purpose of this paper is to give an account, not of the trials and tribulations incident to building the test, but to certain practical uses to which the test has been put.

It should be made plain at the outset that this is no attempt to justify the use of aptitude tests, the value of segregation of students as to ability, or the use of a series of objective tests in measuring accomplishment. It is simply an account of the problems of administration encountered by those in charge of the course in accounting at the University of Oregon with reference at certain points to the use of an aptitude test in attempting to solve some of these problems.

ADMINISTRATION OF SINGLE CLASSES USUALLY SIMPLE

The administration of many university classes is exceedingly simple. The professor in charge of the class is a specialist in his particular field. The class is composed of a relatively small number of individuals who are supposed to be interested in the professor's specialty. In many instances they may be expected to bring a reasonable amount

[1] A. B. Stillman, assistant professor of business administration, University of Oregon. B. A., University of Oregon, 1923. Publication: Joint author of *Interpretive Accounting*, Longmans, Green & Co.

of interest and judgment to bear upon the material presented. In such classes questions of attendance, individual differences in natural aptitude, and even defined standards of accomplishment are of minor importance to the instructor. He is inclined to regard himself as a fountain of information, from which the student may quaff in amounts suitable to his, the student's, individual taste and capacity. The instructor may even defend a certain lack of reliability of examinations and consequent grades on the ground that after all the thing of real value to the student is what he actually carries away with him in the way of new knowledge, inspiration, or ideals, and that the grade received at the end of the term is of little or no real consequence.

In those many courses where the subject is taught to single classes by an instructor whose main interest is in the field involved, and where the class enrollment is made up principally of students whose interest and natural capacities have induced them to explore that particular field of knowledge, the instructor is probably fully justified in devoting most of his thought and energies to enriching the content of the course rather than to the manner of presentation or to objective measures of accomplishment.

GROWTH OF SECTIONED COURSES

Recent years, however, have brought about a situation in our universities where questions of class administration assume considerable importance. There has been a remarkable growth in enrollments in certain courses of general interest. The enrollment in many courses has been artificially stimulated by prescribing them as foundation or "background" courses. There is a marked tendency to reserve any great freedom in electing courses until the junior and senior years, thus forcing students into courses in which they have little real interest. These factors all tend to create a situation where the course must be taught in a number of sections and where the specialist in charge may teach only a limited number of classes. The remaining sections must be taught by graduate assistants or instructors. Questions as to the natural capacity of the student, and as to his interest in the subject may now become problems of great importance. A great variability in the teaching as between instructors may become evident. Differences in individual standards of accomplishment may appear. Differences in objectives may develop according to the preparation and interests of the individual instructors. These things may cause a good deal of dissatisfaction among the students themselves. They also may result in a very unsatisfactory situation from the standpoint of the faculty, particularly in courses designed as preliminary or prerequisite to more advanced work.

ADMINISTRATION PROBLEMS OF SECTIONED COURSES

Problems of administering large sectioned courses fall rather naturally into three groups: Problems having to do with (1) the content of the course itself, (2) the teaching personnel, and (3) the student.

PROBLEM OF COURSE CONTENT

This paper will not attempt to dwell upon the first of these three groups. It is obvious that problems of content will vary with the particular course and with the individual institution. For example, in determining the content of the course in accounting at the University of Oregon, it was decided to put the principal emphasis upon certain interpretive and managerial aspects of accounting rather than upon the acquirement of a bookkeeping technique. Such an objective might not be at all suitable in certain other institutions and the content of the course itself would therefore need to vary with the objective of the institution.

PROBLEMS OF TEACHING PERSONNEL

The problems that have to do with the teaching personnel are, of course, many and at times extremely perplexing. Any effort to secure anything like a standard result with a group of persons naturally as individualistic in their make-up as are college instructors, is bound to be a very difficult and at times an exceedingly annoying problem.

If certain definite objectives could be agreed upon and if a measure of some sort could be applied which would indicate the progress made by individual instructors toward the attainment of that objective, the results of such a measurement would provide a most illuminating and forceful argument in dealing with the instructor. However, in measuring the results accomplished by an instructor, one must first know the quality of material the instructor had to work with. If some measurement of aptitude were had, one might select "pairs" of students of approximately the same aptitude from sections taught by different instructors. A comparison of the accomplishments of the students of similar capacity but receiving instruction under different teachers would be illuminating.

Such a plan was used at the University of Oregon as an aid in evaluating the quality of teaching done in the basic accounting courses during the fall term, 1929. There were 7 instructors who taught at least one section of accounting. Seven groups of 12 students were then found whose aptitude scores indicated that they were approximately equal. These groups had almost the same total aptitude and were almost exactly equal man for man. Each group was selected from the class of a single instructor. These groups were built up as

follows: A student with an aptitude score of approximately 104 was selected for each group, similarly another "pair" with an aptitude of approximately 99 was selected, and so on, until there were 12 pairs selected with aptitude scores ranging from 104, which was a relatively high score, to 68 which was relatively low. In order to make the selection an absolutely impersonal one this pairing was done without access to the accomplishment (criterion) score.

The aptitude scores of the groups is shown below.

Pair	Instructor						
	A	B	C	D	E	F	G
1	2	3	4	5	6	7	8
1	104	104	104	105	105	108	104
2	99	99	99	97	100	99	99
3	97	95	95	95	92	92	95
4	90	89	90	93	92	90	91
5	88	89	89	88	88	89	88
6	86	88	89	88	87	87	88
7	85	85	85	84	84	85	85
8	83	84	84	84	83	85	85
9	81	81	80	83	80	84	81
10	81	80	80	78	79	78	79
11	72	73	72	73	74	71	74
12	68	66	68	72	69	69	68
Total	1,034	1,035	1,035	1,040	1,033	1,037	1,035

The accomplishment score of each student was then determined and tabulated:

Accomplishment scores

Pair	Instructor						
	A	B	C	D	E	F	G
1	2	3	4	5	6	7	8
1	214	203	225	208	250	247	212
2	195	168	209	194	232	181	208
3	228	162	205	176	219	153	199
4	225	194	230	176	186	205	195
5	252	180	184	196	208	231	225
6	224	202	201	192	168	243	205
7	219	141	190	208	176	211	190
8	191	169	191	169	170	210	157
9	197	157	230	171	187	183	190
10	226	188	204	139	178	209	211
11	199	201	126	134	217	166	158
12	185	123	181	187	196	154	210
Total	2,554	2,088	2,376	2,152	2,387	2,393	2,360

The accomplishment scores of the students in each group were taken as the measure of the teachers' efficiency. This may be open to some criticism on the score that the groups are not necessarily equal in application, interest, and other elements which can not be measured by an aptitude test. The answer is that these factors are the things which most test the teacher's skill and that the factors

which are not susceptible to the influence of the teacher are probably a minor consideration. Instructors ranked as follows:

A with a score of 2,554	I
F with a score of 2,393 E with a score of 2,387 C with a score of 2,376 G with a score of 2,360	II
D with a score of 2,152 B with a score of 2,088	III

The typical accomplishment of careful and efficient teaching is probably represented by the scores in Group II.

A similar plan had been used in 1928, with a good deal the same sort of grouping as the result. In the fall of 1930, instead of picking pairs, the entire section was used as the basis for comparison, by reducing the aptitudes and accomplishments to standard scores. The results were very similar, however, with the tendency to show that there were one or two instructors who did not measure up to quite the same standard of result as set by the main group.

Such comparisons of achievement and aptitude are very helpful to those in charge of the course. One should state, however, that these comparisons are never used as the sole basis for judgment of the teacher's efficiency but rather as important corroborative evidence.

Two particular results have been noted from this use of the aptitude test. First, the evidence seems to support the theory that it is the careful and painstaking teacher who produces the most consistent and efficient results, rather than the more brilliant but somewhat careless type. Second, there is a tendency for the teachers to reach about the same plane of accomplishment when they are aware of the results of such a measure. This last may be partially due to the opportunity it gives the administrator to put pressure in the right place. At any rate, the use of the aptitude test in this way has tended to equalize many of the usual differences in teaching.

PROBLEMS HAVING TO DO WITH STUDENTS

The problems that have to do with the students enrolled may be listed as follows:

1. Should the student be in the course at all—are his interests and capacities such as fit him for this type of work?
2. Are there any considerable numbers of students whose preparation or backgrounds are such that they would do better work if sectioned by themselves?
3. Can one discover groups of varying abilities, and can such groups be segregated so as to make possible the development of methods and the adjustment of course content to better meet their needs?

SHOULD THE STUDENT BE IN THE COURSE?

One of the most difficult situations which arise in connection with large sectioned courses concerns the student who has no natural aptitude for work in that particular field. The problem of treating such a person fairly and justly becomes much aggravated when the course is a prescribed course which the student must take no matter how distasteful it may be. Our usual treatment of such cases is to make the student secure a grade in open competition with scores of persons naturally better fitted to do the work of the course. If he fails to measure up to the standard set by this larger group of interested and more capable persons, we give him a grade of F. The whole proceeding is about as just as requiring a one-legged man to run a foot race in competition with normal men, and shooting him if he comes in last. The use of the aptitude test reveals that many cases of so-called indifference and laziness are really cases of a definite lack of ability. Unfortunately, although the use of an aptitude test may reveal the reason for failure under such circumstances, it does not suggest a remedy. However, the aptitude test does serve to reveal such a situation and challenges the conscientious administrator to work out some remedy.

SEGREGATION AS TO BACKGROUND

Even with those students whose aptitudes are somewhat higher one finds a wide variation of background and interests. Such variations are not likely to be evidenced by the aptitude test. In the course in accounting at the University of Oregon there are about 50 women in the group of 450 students enrolled. It was found that 3 or 4 women in a group of 30 or 35 men apparently were at a disadvantage. A separate section for women was arranged. The work of the young women seemed to improve somewhat, their interest seemed to be materially better, and the problem of teaching was simplified. This experience has suggested some interesting fields for investigation.

SEGREGATION AS TO ABILITY

The most obvious use of the aptitude test would seem to be to use the information thus found as a basis for segregating students into groups of similar ability, providing, of course, that such a segregation would result in a proper adjustment of the material given to the needs of the respective groups.

Frankly, the administrators of the course in accounting at the University of Oregon have never found themselves ready to rely upon the aptitude test as the sole basis for such a segregation. Rather, for a period of three or four years, students have been sectioned in heterogeneous groups during the fall term. At the beginning of the winter term and again at the beginning of the spring term they are resec-

tioned on the basis of the instructors' grades. The development of a more careful and uniform system of grading has tended to make the grades given by individual instructors much more equable and just.

APTITUDE SCORES AND THE INDIVIDUAL STUDENT

There are many other ways in which the aptitude score has been used. A particularly important one is in connection with the personal contact between the instructor and the student. In this case we do not rely entirely upon the special aptitude score but take into account the psychological rating given by the personnel bureau to all entering students together with a comparable score of high-school accomplishment, also compiled by the personnel bureau. Thus, if a student has a high score in his general ability and a relatively low score in his high-school record, we have a fairly accurate idea as to his study habits. If his natural aptitude in accounting, as revealed by our test is low, we must realize we have a certain problem in teaching; he must be watched closely and made to keep up on his work. If rather exceptional natural ability in accounting is indicated, the problem will be to stimulate his interest and correct his habits of study.

APTITUDE TEST AS A BASIS FOR JUDGING ACCOMPLISHMENT

At the present time the special thought of those in charge of the course has shifted away from further perfecting of the aptitude test to the development of new methods of presenting the material. The present plan of segregating students as to ability has one fundamental weakness. This plan has been to have all students cover the same ground, but to so arrange the matter as to permit the more capable persons to go into each point more intensively. The way this is working out is that the more able student does more work than the less able. So far as the practical necessities of the situation are concerned, it ought to be the less apt student who does the most work. He is the one who most needs it. To correct such a situation may require some very radical and unusual practices. We are considering at present the setting up of an experimental group of say 100 students. The work of these students would be set up in "budgets," in many respects not unlike the contract system used in the secondary schools. Periods for supervised study would be arranged. The student would be permitted to carry on the work as rapidly or as slowly as his natural ability would permit, taking frequent objective quizzes and being required to repeat any portion of the budget which his quizzes indicate has not been thoroughly mastered. The remaining sections would be operated much as at present. Here again the aptitude test will serve a most useful end in giving a basis from which to measure the relative efficiency of the two methods of instruction.

SUMMARY

The matters set forth in this paper should be regarded as the account of an experiment which has been carried on at the University of Oregon for a period of several years. It is an expression of the views of those in charge of this work and is decidedly not to be interpreted as any effort to suggest that these views are applicable to any course or to every school. Those administering the course have found the development of an aptitude test exceedingly helpful in several ways. It has been an aid to them in measuring the effectiveness of the various teachers; it has helped in securing greater uniformity in objectives and in teaching method; it has acted as a reenforcement to the instructors' judgment in segregating students on the basis of grades; and it has served as a criterion in judging the accuracy of our tests. Perhaps most important of all has been its effectiveness as a guide to the problem of the individual student, particularly when used in conjunction with the ratings for general scholastic ability and for high-school accomplishment.

A most important though more indirect result in this particular instance has been the way in which the effort to perfect such a test has awakened those in charge to some of the really pressing problems of administering a large sectioned course.

ESTABLISHING A STUDENT MENTAL HYGIENE CLINIC

OTHNIEL R. CHAMBERS [1]

The need, in an effective personnel program, of such a service as only a mental hygiene clinic could supply has been increasingly recognized since the days of the World War. Since Morrison and Diehl's study (4)[2] in 1924 the percentage of serious cases has been fairly well established at a minimum of 8 or 10 per cent. Doctors Riggs and Terhune (5) say 10 per cent in 1928.

That it is not the poor student alone who is in need of aid is stressed by such studies as the anonymous study (1) published in 1921 showing that some two-thirds of the Phi Beta Kappa graduates of one institution have shown signs of neuroticism or psychoneurosis.

The chief contention has arisen as to the method of establishing the needed mental hygiene service. Certain critical factors must be considered in deciding upon the mode of attack. Some of these include (*a*) the groups in need of aid, (*b*) the scholastic ratings of these groups, (*c*) the development of wholesome campus attitudes toward the service, (*d*) the content of the course to be given as the basis of the mental hygiene service, (*e*) the attitude of the faculty toward the service, (*f*) the securing of maximal returns on a minimal investment, (*g*) the consequent use of all personnel and equipment already available upon the campus, (*h*) securing the public's interest in the venture in order to (1) educate the public and advance the mental hygiene program in general, and (2) secure outside support for the institutional program.

The individuals needing aid may be grouped in three groups: (*a*) The large group of students who could secure sufficient aid from class instruction in mental hygiene; (*b*) a smaller group whose problems are of such a nature as to require more or less personal aid, analysis, etc.; and (*c*) a small group whose mental condition is too serious to justify care by the institution and who will have to be withdrawn and placed in the hands of psychiatrists with hospital facilities.

---

[1] Othniel R. Chambers, professor of psychology, Oregon State College. A. B., M. A., Indiana University, 1922; Ph. D., Ohio State University, 1926. He was formerly connected with the departments of psychology at Ohio State University and the University of Texas. Publications: "First Revision of a Group Scale for Investigating the Emotions," *Journal of Applied Psychology*, 37 : 97-104; "Relation of Intelligence and School Training to Observational Learning," Bulletin, Extension Division, Indiana University, 7:12; "Character Trait Tests and Prognosis of College Achievement," *Journal of Abnormal and Social Psychology*, 20:3:309-311; "Method of Measuring the Emotional Maturity of Children," *Pedagogical Seminar-Journal of Genetic Psychology*, 33:4:637-647; "Measurement of Personality Traits," *Research Adventures in University Teaching*, chapter 9, Public School Publishing Co.

[2] Numbers in parentheses refer to "Bibliography," p. 1160.

As you note, these groups are of rapidly decreasing size, of rapidly decreasing importance to the institution, and to the State at large, and of very markedly decreasing hopefulness from the point of view of prognosis.

The first group is outside the 10 per cent which has always been stressed, and its exact or even approximate size is unknown. The second and third groups make up that much-discussed 10 per cent referred to in the opening paragraph.

The second major question concerns the year of school in which the greatest need lies. It is rather evident that the problems will be more marked in transition periods and hence the freshman year would be expected to yield considerable maladjustment. This is actually found to be true. It would seem then that the place to begin is in the freshman year.

The mode of attack is rather vital. Publicity is not desired as wrong impressions are nearly certain to gain credence and the work be looked upon as work with "nuts"—the abnormal.

As mental hygiene stresses prevention and much the larger group needs only that service which can be rendered by class instruction, it is advisable to enter a wedge by the establishment of a freshman course in mental hygiene. This should, the writer believes, be made an elective course. The reason for this is that the attitude requisite for mental hygiene aid and mental therapy is not secured by forcing that aid upon the individual to be helped. Moreover, the maladjusted in college are much more likely to be hostile to authority than are the well adjusted. The content of such a course has been very adequately outlined by Doctor Blanton (2) and quoted by Bohannon (3). Of necessity that content must be adjusted to the local situation. An instance may be given. At our institution three orientation courses are required of large numbers of students. One of these courses is "How to study." This topic, then, which appears in the Blanton-Bohannon outlines had to be dropped. Local situations as regards the social significance attached to fraternity and sorority membership, the extremes between rural and urban culture, etc., will also alter course content or determine illustrative material.

Certain it is that the course should not be a course in abnormal psychology and should not induce a morbid attitude on the part of the students. The appearance of, the cause of, the effect on efficiency and mental health, and removal of such things as exaggerated emotions, day dreaming, homesickness, inferiority feelings, rationalization, phobias, unhealthy attachments, compensations, fatigue (mental and physical, etc.), should be stressed—not from the point of view of the abnormal but from that of their appearance—even in you and me.

The development by each student late in the course of an autobiographical case study, and the giving of a number of tests in the

course and their grading by the student gives the student an estimate of his own needs and points out to the instructor those needing personal aid.

The granting of clinical aid to those desiring it—its use being purely optional—brings in almost all those students needing aid.

In the writer's experience 10 per cent of the classes came in during the early days of the course. Now more and more people needing the course are electing it. Students advise friends to take the course. Deans, instructors, and student advisors recommend it. Eventually the course becomes a more or less effective selective agent taking in a larger and larger percentage of those needing aid. Last term 17 per cent of those taking the course came to the clinic for personal aid.

The Thurstone and Thurstone psychoneurotic inventory (6) has been found extremely effective for giving in the class. Boys making scores over 67 and girls making over 75 are, in more than 98 per cent of the cases, in real need of personal aid. More than 85 per cent of boys making over 58 and girls making over 65 are in need of such aid. This test, while it does point out certain types of cases, does not point out others. A low score on this test is no guarantee of emotional stability.

The Thurstone and Thurstone psychoneurotic inventory should not be given until rapport is established with the student, nor should interviews be proffered before fairly late in the course, for the same reason.

This teaching of mental hygiene is not a mere incidental in a program of this type. It is the very groundwork, acting as a preventive, as a sales force to the student body and to the faculty, and bringing in cases for clinical study by an "endless chain" method. Its cost is light.

In the establishment of the clinic, advantage should be taken of all available personnel and equipment. The instructor in mental hygiene, and the case worker—a member of your psychology staff—should bear no official administrative position such as deanship, assistant deanship, etc., if he is to secure absolute confidence and also full information. Moreover he should have a broad toleration. While the case worker should, in the writer's estimation, have a religious faith, he should not be very closely identified with any sectarian organization as such identification will tend to prevent certain confidences. No case should ever be handled by a person whom the student's problem upsets emotionally.

It would be most unfortunate to have an out-and-out Freudian or disciple of any one school in charge of the clinical work.

The use of the staff perhaps can well be illustrated by our situation at Oregon State College. Work in speech adjustment is shunted to Professor Wells, of the speech department; reading difficulties are

turned over to Doctor Parr; vocational guidance problems go to Professor Salser; physical examinations are made with especial care by the physicians of the health service; aid in religious problems is secured from Reverend Warrington or from the pastors of local churches.

The establishment of such clinical facilities should be quiet, non-advertised, nonspectacular in growth, and founded on mental-hygiene instruction in the freshman year.

BIBLIOGRAPHY

(1) Anonymous. Mental Hygiene and the College Student Twenty Years After. Mental Hygiene, 5:736–740, October, 1921.

(2) Blanton, Smiley, M. D. A Mental Hygiene Program for Colleges. Mental Hygiene, 9:478–488, July, 1925.

(3) Bohannon, Charles D. Mental Hygiene from the Standpoint of College Administration. Annals of the American Academy of Political and Social Science, 149: part 3:86–101, May, 1930.

(4) Morrison, Angus W., M. D., *and* Diehl, H. S., M. D. Some Studies on Mental Hygiene Needs of Freshman University Students. Journal American Medical Association, 53:1666–72, Nov. 22, 1924.

(5) Riggs, Austin Fox, *and* Terhune, William B. The Mental Health of College Women. Mental Hygiene, 9:261–272, April, 1925.

(6) Thurstone, L. L., *and* Thurstone, Thelma Gwinn. A Neurotic Inventory. Journal of Social Psychology, 1:3–30.

TEACHER-APTITUDE TESTS AND TEACHER SELECTION

NELSON L. BOSSING [1]

I. THE NEED OF DISCOVERING BETTER SELECTIVE DEVICES FOR TEACHERS

In the field of education we have become increasingly conscious of the need of a better type of selective device than the usual ones employed if we would choose and direct wisely those who should enter the teaching profession. When teachers were few our chief problem was to find a sufficient number of teachers to man our schools. That day is past. We are now confronted with a numerical oversupply of teachers. It is now a question of the selection of the best available for training and placement.

According to the Bureau of Education Bulletin, 1929, No. 17, on Teacher Training:

> The number of students enrolled in all types of institutions which train teachers is more than one-half million. This is more than 400 per cent greater than the number undergoing training two decades ago. During the same period the number of teaching positions has increased by approximately 35 per cent.

Again quoting from another Bureau of Education Bulletin, 1929, No. 14, which discusses the number of teachers now in training for public-school positions throughout the United States, we find this startling conclusion:

> It is safe to assume that these institutions (colleges and universities) are interested primarily in the preparation of teachers for high-school positions. If so, we face the probability that at this time there is a student in training for every high-school teacher position.

Professor Miller, of the school of education, Columbia University, in the June, 1929, issue of the High School Teacher estimates that we have an oversupply of 150,000 teachers.

This, of course, is not the entire story. Another consideration that makes urgent better selective devices is the fact that among those in preparation for or now teaching, we have a large number of misfits, people who have met the standards of preparation required for certification and yet who find themselves temperamentally or through other deficiencies incapable of successful work in the profession.

---

[1] Nelson L. Bossing, professor of education and director of supervision, University of Oregon. A. B., Kansas Wesleyan University, 1917; M. A., Northwestern University, 1922; Ph. D., University of Chicago, 1926. He was formerly head of the department of education and director of summer sessions at Simpson College. Publication: "History of Educational Legislation in Ohio from 1850 to 1925," F. J. Heer Publishing Co., Columbus, Ohio, 1931.

A third factor which gives emphasis to the matter of selection is the fact that professional standards in education have become quite rigid and indications are that the standards for the profession of education will soon approximate the level of some of the recognized professions, in the quantity of training required. For example, we now require four years of training for the high-school teachers in the State of Oregon, whereas not a few years ago graduation from high school was sufficient academic preparation with which to enter high-school teaching positions. Our two neighboring States—Washington and California—have each entered upon a 5-year plan of preparation for those who are to be certified as high-school teachers. Obviously everything that humanly can be done should be done to safeguard both the prospective teachers and the profession against the preparation of those who lack the necessary qualifications for professional success. It should be possible to prevent the unpromising candidate from entering upon such an extensive course of training.

Still another factor which looms large in these days of economic depression is that of teacher-training costs. According to the United States Bureau of Education Statistical Circular, No. 11, on Per Capita Costs in Teacher-Training Institutions, 1927–28, we find approximately $300 the average cost reported for a year's training of prospective teachers, with wide variation for individual institutions of from $194.80 to $439.67.

Against this general need which flows from the increasing demands of the profession of teaching we must face frankly the inadequacy of past and present methods by which we have and do attempt to determine who are and who are not good teaching risks.

II. ATTEMPTS TO DETERMINE TRAITS OR FACTORS THAT RELATE TO TEACHING SUCCESS

Among the earlier efforts to evaluate teachers was that of the rating scale. The use of the rating scale has reflected change of emphasis over a period of years. The first scales were used by school superintendents and school administrators to determine the merits of teachers within their employ either for promotion or elimination from the system. Later the scale was looked upon by leaders in the profession as a device by which to improve the teaching power of the teacher herself. It is only within recent years that attention has been focused upon the desirability of determining traits or factors that might have prognostic value in determining who should and who should not become teachers. The attempt, therefore, to determine specifically and as far as possible objectively the presence of measurable traits or factors of teaching success or aptitude is relatively of recent date. Dr. F. B. Knight (6) [2] in his doctor's study at

[2] Numbers in parentheses refer to "Bibliography," p. 138.

Columbia University recognized three studies of exceptional value on this subject prior to that date. They were J. L. Meriam's study on Normal School Education and Efficiency in Teaching, carried on at Columbia University and published in 1906 (11); The Development of a Grade Scale, by Dr. Edward C. Elliott, in 1910; in which he developed a tentative scheme for the measurement of teaching efficiency, later revised; and a third study was an extensive research by A. C. Boyce and published under the general title "Methods for Measuring Teachers' Efficiency." (2) Prior to 1925 but some half dozen studies of significant value seem to have been undertaken. In addition to those previously mentioned, Dr. G. T. Somers' Pedagogical Prognosis: Predicting the Success of Prospective Teachers (16) another research study for a doctor's degree at Columbia University in 1923, and the monograph of Dr. F. L. Whitney on The Prediction of Teaching Success, published in 1924, comprised the battery of studies considered of value (19). Since that date some two dozen studies of varying degrees of merit on this subject have been made and reported in periodical literature or monographic form.

For some time it has been generally assumed that scholarship was a quality that had large significance as a means of predicting future teaching success. Unfortunately a number of studies that have been made do not give us cause for great confidence in this factor as an instrument of prediction. For example, the results reported in a number of studies such as that of S. A. Hamrin's (5) show a correlation of only 0.05 between school marks of teachers in training and the later ratings of these same teachers by superintendents in the field. Dr. F. B. Knight (7) in the study previously referred to found a correlation of only 0.153 between ability to teach and scholarship. Doctor Whitney in his study found a correlation between academic marks and teaching success after graduation of but 0.073 (20). Roy R. Ullman in a study of the prediction of teaching success carried on for his master's degree at the University of Michigan (18) found a correlation between teaching success in the field and general scholarship of 0.30.

G. P. Cahoon in an article in the May, 1930, issue of the University High School Journal (3), reports a correlation of 0.065 ± 0.05 between academic grades for one group of students correlated against practice teaching success, and for another group of the same year a correlation of 0.27 ± 0.05. He, therefore, concludes "that there is no relation between success in college as indicated by general college marks and success in practice teaching."

He further states "that the factors of success in each of the two situations are somewhat different. In college the emphasis is largely upon achievement in subject matter while in the secondary school it is upon achievement in instructing pupils with subject matter more as a means to an end."

This is a reasonably fair sampling, I think, of the findings of some of the better studies that relate to the factor of scholarship as an element of prediction of teaching success. This should be said, however, in behalf of the findings of Mr. Ullman. His correlations with the 12 factors studied are generally low.

Another factor that has been used as a basis of determining the availability of teaching candidates has been that of their professional record. Again Mr. Ullman (18) found a correlation of 0.30 between professional courses in education and later teaching success. Doctor Whitney (20) found a much higher relationship between teaching success and professional marks than with academic marks. He found a correlation of 0.143 for professional marks.

Another factor frequently considered possible as a means of predicting teaching success is that of intelligence. However, there seems to be general unanimity on the part of those who have made worthwhile studies at this point that there is not sufficient relationship between intelligence and teaching success to warrant confidence in intelligence as a predictive trait. Knight (8) used the early Thorndike college entrance examination as a measure of the intelligence of high-school teachers and found a correlation between intelligence and teaching success as estimated by teachers and supervisors of 0.41. Somers (17) in his research study for his doctor's degree found that the correlation of intelligence as measured by mental tests and success in teaching gave a coefficient relationship of 0.43. Whitney on the other hand found a correlation between teaching success after graduation with intelligence as measured by intelligence tests of 0.025 (20). Possibly the best of the later studies was that made by W. H. Pyle (15). He correlated scores on the Detroit Advanced Intelligence test with grades made in practice teaching and found a correlation of only 0.153 ± 0.035. Again intelligence scores were correlated with the teaching of these students after the first year in the public schools. Ninety-nine cases used as a basis of the study gave a correlation of 0.034 and a correlation for the second-year teaching in public schools of 0.023. In both situations the probable error was 0.066. Pyle, therefore, concludes that intelligence of students has no considerable value in predicting the later teaching success as judged by the criterion of principal's judgments of teaching success. Similarly Ullman (18), by the use of the Brown psychology test, found a correlation of but 0.15.

Again we have frequently utilized cadet teaching grades as predictive of later teaching success and the evidence would seem to suggest that our highest correlations are to be found between the practice teaching of the prospective teacher and later success in the field. Meriam found a correlation of 0.443 (12) between practice teaching during the normal school training and teaching success after gradua-

tion. Whitney (20) found a correlation between teaching success after graduation with student teaching of 0.238. Ullman (18) in the study previously referred to found a correlation of 0.36. W. W. Ludeman in a study made at Ohio State University reports a correlation of 0.63 between practice teaching and later teaching success (10). These correlations while not high do suggest the value of cadet teaching grades as indexes of later success.

It is evident, however, that cadet teaching, valuable as it is as an index of future teaching success, does not meet the need of a predictive device by which to determine the fitness of entering candidates into the teacher departments of our teachers colleges and universities because cadet teaching is usually the last thing which the student does before entering his professional career.

It is quite evident from a careful study of investigations thus far pursued that a careful analysis of our criterion has not been a matter of concern by students of this subject. Most investigators have assumed the value of their criterion without further investigation. To the writer's knowledge only one or two studies have been made in which the criterion has been subjected to careful appraisal. Yet he finds himself in full agreement with Doctor Jacobs that in the absence of better standards the opinion of competent judges must be accepted. Doctor Jacobs (21) succinctly puts his case thus:

> From the work of these investigators and from readings in related fields, e. g., personnel management, there emerged the conviction that the most reliable criterion at the present time with respect to teacher effectiveness is the conscientious and deliberate opinion of competent judges.
>
> The theory underlying this conclusion is that where objective measures in terms of amount are not fully applicable to determining the degree to which a given characteristic or group of characteristics is present in a certain situation, then the opinion of competent judges must be resorted to either wholly or in part. This hypothesis is the basis for practically all social measurements. It must be granted that, as in the case of this study, while it is possible to establish the reliability of personal judgments, there is no way of proving their validity. It is because of this difficulty to prove validity that attempts are constantly being made to find a truly objective method of determining teacher effectiveness. The greatest difficulty here, and this is almost universally overlooked by the investigations in the field of teacher rating, is that the product of the teacher's effort really comes to fruition not in a week, nor in a month, nor yet in a year. The fact is some 10 to 20 years must pass before the fruition is attained.
>
> It is true that in the absence of means for measuring the actual product, measurements that are predictive of ultimate effectiveness must be resorted to. But, if we do resort to such measurement as a method of estimate, we must be sure that what we measure is valid as a criterion of estimating. And since being sure is a matter of opinion, we are led to the comment that there are certain conditions under which opinion must be accepted as fact.
>
> Until the time arrives, then, when an objective method for determining the comparative effectiveness of different teachers has proven both valid and reliable, the subjective opinion of competent judges must continue to serve the purpose.

However, if we are to find adequate leads for the discovery of traits or factors of success it is necessary to subject the criterion to careful analysis for what it may yield.

During this year we have undertaken to determine the value of the criterion by which teachers are adjudged successes or failures in the field for the possible purposes of discovering predictive elements that might lead to the development of a predictive test. An attempt has been made, therefore, to discover what elements make up the total value of the criterion. For this purpose we undertook a study of the judgments of superintendents and principals who have rated our graduates from the school of education while at work in the field.

III. A STUDY OF THE CRITERION OF TEACHING SUCCESS USED

The general plan was as follows: For a number of years past the teacher placement bureau of the University of Oregon annually has sent to the superintendents and principals of secondary schools uniform rating blanks upon which to check the success and progress of teachers who are graduated from the University of Oregon. As a result of this policy over a period of years we had two or more ratings for a large number of teachers now in the field. The study began with the rating of 248 teachers who had taught two or more years and for whom we had received rating forms from their superintendents and principals.

Because of the lack of complete and accurate data the number of teachers who could be used for this study was reduced to 165 cases. These 165 teachers represent 84 school systems which range in size from 2-teacher to 54-teacher high schools. For each of these teachers we had complete ratings representing two different years. In 93 cases the teachers were rated by different judges and in 72 cases the teachers were rated by the same principal or superintendent. In every case a year elapses between the ratings. In 47 cases the teachers were working in different school systems when they were judged the second time. Since it has been our custom to send a rating form for each teacher annually sometime in January, the judges had no knowledge of previous ratings given to the teacher. The academic and professional grade averages for these 165 teachers used for supplementary purposes in the study were taken from the registrar's office of the University of Oregon.

Treatment of data.—Finding reliability of criterion in the treatment of the data secured, we began the study by an attempt to establish the reliability of our criterion. As mentioned before, the criterion is based upon the judgments of superintendents and principals who supervised the work of the teachers judged for one or more years. The ratings were recorded on regular forms sent out by the placement

bureau and returned to the bureau where they were filed. The rating form has 13 items to be evaluated by the principal. However, because certain items on this rating form were not applicable to the study we eliminated items Nos. 10, 11, and 13. Item No. 10 requested a judgment of the weakest point and No. 11 of the strongest point in the repertoire of the teacher's equipment. No. 13 was a general invitation for comments not otherwise suggested on the rating form. The 10 items made the basis for our study were as follows: (1) Ability as instructor; (2) success in discipline; (3) industry; (4) character; (5) personality; (6) personal appearance; (7) health; (8) loyalty and cooperation; (9) attitude toward community; (12) general rank.

For the first nine items the following words were used to rate the degree of success in each, namely: Very best, which was given a value of 6 points; Excellent, 5 points; Good, 4 points; Medium, 3 points; Inferior, 2 points; Failure, 1 point. The "general rank" was item No. 12 on the score card, followed by these 6 adjectives or phrases, one of which the judge underscores to indicate his general opinion of the teacher's success—Distinctly superior, very good, good, average, slightly below average, poor. In the determination of the reliability of these judgments the correlation technique was employed. In other words, the ratings of the teachers received the first year were correlated with the ratings received the second year.

The correlation of item No. 12, "general rank," was first secured independent of all other items on the rating form: The correlation between the two sets of judges gave a correlation coefficient of 0.828 ± 0.020. A total of the judges' ratings on the first nine items was next studied, for which a correlation coefficient of 0.052 ± 0.131 was found. This would suggest very little agreement on the part of the judges when they consider the different traits on this rating form but indicates considerable agreement when they consider the above "general rank" of the teachers.

Following these two attempts at a general or composite rating a study was made to determine the part each of the nine items on the rating form played in the total ranking the judges gave the teacher. The weighted value of each of these nine items was determined by the beta regression equation, using partial multiple correlation technique. The values found are as follows:

1. Ability as instructor	0.779	6. Personal appearance	0.191
2. Success in discipline	.289	7. Health	.299
3. Industry	.210	8. Loyalty and cooperation	.191
4. Character	.040	9. Attitude toward community	.063
5. Personality	.071		

From these data it is evident that "ability as instructor" is almost as important in the estimate of the judges as the other eight items taken together. It further suggests that "ability as an instructor" has much the same implication for the judges as item No. 12, "general rank." This point is further emphasized when we note the correlation coefficient between item No. 1 of one set of ratings and item No. 1 for the rating given the following year or a year later. This fact is further borne out by a comparison of the correlation coefficients between the several items on the one set of ratings compared with similar items on the second set. The comparisons for the two years are as follows:

Ability as instructor	$r_{11} = 0.792 \pm 0.114$
Success in discipline	$r_{22} = .241 \pm .101$
Industry	$r_{33} = .021$
Character	$r_{44} = .201$
Personality	$r_{55} = .201$
Personal appearance	$r_{66} = .301$
Health	$r_{77} = .492$
Loyalty or cooperation	$r_{88} = .002$
Attitude toward community	$r_{99} = .023$

The conclusion to be drawn here is the same conclusion that is drawn by Mr. Knight in his study Qualities Related to Success in Teaching, where he found that there was no evidence of ability on the part of judges to analyze and weigh the factors that entered into the general judgment of teaching ability. His conclusions, however, were drawn from almost diametrically opposite results, because most of his correlations on individual factors were high while the relationship in certain instances could not be justified on any basis of rationality. For example, he raises the question as to what rhyme or reason there is in a correlation of voice with intellect of 0.682 or a correlation of voice with accuracy of 0.628. He concludes, therefore, that there is no rational basis for the relationship between factors as found in his study, but that his judges were influenced by the halo effect of their general judgments upon specific factors (9).

As a further check upon the criterion and its possible significance for our judgments of prospective teachers we sent to superintendents and principals of the 165 cases studied a somewhat different rating form which is used by the school of education in determining the factors of success of our cadet teachers. This special form is filed with the director of supervision at the close of the teaching period of each cadet along with a grade for the cadet which is represented by the general rating given the cadet on this rating sheet. This rating sheet was sent out to the superintendents and principals concerned some six weeks after the 1931 appointment bureau rating form had been returned to the appointment bureau.

This form contains four major sections designated: (1) Personality equipment. (2) Social and professional equipment. (3) School management. (4) Teaching skill. Under these four classifications occur 36 items upon which a judgment was asked. The correlation was made between 100 of these forms that had been returned by the fifth of April with the similar rating forms on file in the appointment bureau for these same teachers while cadets. The following coefficients of correlation were obtained for the four major items: (1) Personality equipment, $r_{11}=0.506$. (2) Social and professional equipment, $r_{22}=0.615$. (3) School management, $r_{33}=0.440$. (4) Teaching skill, $r_{44}=0.184$.

The sum of the correlations for the four major items on the scale gave a total correlation value of 0.476. These figures indicate first, that while there is a higher degree of interrelation between the four major items on the efficiency record forms used with our cadet teachers and judgments in the field, the value for the total rating of these items of 0.476 does not give as high results as the correlation on the appointment bureau rating form where we secured 0.520 as the total ratings of the items studied. We are, therefore, prone to the same general conclusion that was reached in the study of appointment bureau rating forms, namely, that a general judgment is more susceptible of agreement between judges than is an attempt to correlate factors that make up the total judgment. Further, we conclude that a simple rating device seems to assure greater predictive accuracy than a more complicated one.

In order that we might further check against the possible values of this study as a basis of comparison with other studies we secured correlations between the criterion and (1) cadet teaching grades, (2) professional educational grades, and (3) all academic grades not including grades in professional education courses. The correlations of these grades are as follows: $r_{1c}=0.687\pm0.072$; $r_{2c}=0.188\pm0.056$; $r_{3c}=0.172\pm0.088$. These results are in agreement with the findings of other investigators reported in the first part of this paper.

The results of our study, therefore, would indicate that considerable confidence may be placed in the cadet teaching grades for prediction of success in later teaching but that grades in professional subjects or academic subjects for purposes of prediction are of very doubtful value.

Because most of the studies reported are based upon the use of comparatively few cases it seemed desirable to experiment from samples—our 165 cases to determine the reliability of the criterion. Using the same ratio of different judges against the same judge as existed for the total of 165 cases, 57 cases were taken at random which resulted in a correlation coefficient of 0.378 ± 0.092. Another sample taken on the same basis but including but 20 cases resulted in a corre-

lation of 0.90 ± 0.027. These checks serve to give considerable pause to the dependence to be placed in correlations where the total number of cases is small. It is characteristic of most of the studies made on this topic that the number of cases is relatively small. For instance, Doctor Morris in her doctor's study on Personality Traits and Success in Teaching used but 60 cases to validate her study. Ullman in his recent study referred to above used but 116 cases. Boardman in his doctor's study of Professional Tests as Measures of Teaching Efficiency in High School (1) had complete data for only 88 teachers who participated in the study. Knight in his doctor's study previously referred to used three school systems which involved a distribution of teachers as follows:

School A.—Elementary teachers, 53; high-school teachers, 1[illegible]
School B.—Elementary teachers, 35; high-school teachers, 13.
School C.—Elementary teachers, 30; high-school teachers, 10.

However, the number of useful ratings he was able to secure totaled but 97. Part of his technique involved the study of correlations in each school system. The result was that some of his conclusions were based on extremely meager data. In a survey of the literature thus far but few studies have used a larger number of cases than employed here. The conclusions therefore that we would reach as far as this study is concerned are:

1. Very few writers in this field have established the reliability of their criterion.
2. Most studies are based on too small a sampling.
3. General ratings by judges are more reliable than the judgment of individual factors that go to make up the general rating.
4. Rating devices with few items to be scored will give a higher reliability than in the case of rating forms with a large number of items.
5. The predictive value of cadet teaching grades at the present time seems to offer a better basis for the determination of future teaching success than any other criterion.
6. Apparently little confidence can be placed in academic grades, professional grades, or intelligence ratings as a prediction of teaching success.
7. A corollary conclusion to the above would be that an analysis of the criterion used in this study does not offer much suggestion as to what particular elements should enter into the formation of teaching aptitude tests.

IV. APTITUDE TESTS

During the last few years with the increasing evidence that no one specific factor might be utilized as a means of predicting teaching success a number of investigations have been made in an attempt to

discover a battery of traits or factors which taken as a composite might serve as a practical instrument of prediction. The results of some of those studies have found expression in the development of aptitude tests which attempt to give a general prognostic index of the individual's success as a teacher. Among these tests may be distinguished roughly four types. The first more or less general type may be represented by the aptitude test for elementary and high-school teachers, by Bathurst, Knight, Rugh, and Telford. This test consists of six subtests under the following titles: (1) A professional judgment test. (2) A test over the theory and practice of teaching. (3) A test covering reading comprehension. (4) A test of social information. (5) A test of school and class management. (6) A test of professional information.

The validity of the test is determined by the correlation of "the scores on this test with the judgments of superintendents and supervisors who had observed the teaching efficiency of each teacher judged for at least nearly one school year."

This general procedure was followed both for the elementary school teachers and the high-school teachers. A multiple correlation between the judgments and the scores for elementary teachers on the various tests was found to be 0.414, and for high-school teachers, 0.54. There is evidence that this test has some value as an instrument of prognosis. However, until the reliability of the test has been further studied, it should be used with caution.

A slightly different test but of somewhat general character is the Stanford educational aptitude test devised by Milton B. Jensen. The test attempts to be more specific than the aptitude tests for elementary and high-school teachers, since it claims to predict the candidates' fitness for specific aspects of professional education, namely (1) a combination of teaching-research, (2) a combination of research-administration, and (3) a combination of teaching-administration. The test is a rather complex one to score and the test itself is cast in a form of situations to which the candidate must give a judgment. These situations are classed under three general headings, or rather, form three separate tests:

1. Position preference ratings in which the prospective teacher is asked to indicate a preference for type positions that might be available.

2. Discipline case problems in which certain judgments are required in the solution of the problem set up.

3. This test centers around high-school activities and again asks for a judgment in terms of certain general situations that appear in the normal activities of the school, particularly as it effects com-

munity relationship. The author describes the plan by which the validity of the test was established as follows:

> Each item of the test has been carefully weighted in such a manner as to give its greatest possible contribution to each of the scales. Members of the criterion groups (205 men in number) were selected on the basis of ratings by from one to seven judges. Maximum values were obtained from these ratings by means of statistical procedure involving reliabilities and variabilities of the various judges' ratings and the number of judges rating each individual. From these weighted ratings cases were selected in such a manner as to give the best categorical selections with the smallest probable errors. Experience with the test indicates that coefficients of correlation of from 0.80 to 0.90 may reasonably be expected between test scores and ratings such as were used in selecting the criterion groups. It was found that differences measured by the test are independent of age, sex, professional training, and professional experience, and that they are closely associated with self-ratings by individuals who have had extensive professional training and experience. The relationships above mentioned are shown by the correlations of the tests.

This test along with the test by Bathurst, Knight, Rugh, and Telford is considered by Max E. Engelhart in his appraisal of standardized tests for students of education as the two outstanding teaching aptitude tests now available (4).

Another general aptitude test which has received considerable attention is known as the George Washington University teaching aptitude test. This test is divided into five parts: (1) Judgments in teaching, (2) reasoning and information concerning school problems, (3) comprehension and retention, (4) observation and recall, and (5) recognition of mental states from facial expression. However, this test has been subjected to rigorous and experimental procedures by Anna R. Markt, of the National Kindergarten and Elementary College of Chicago, and Prof. A. R. Gilliland, of Northwestern University. In brief, they gave the test to a group of 145 freshmen girls in the National Kindergarten and Elementary College with no teaching experience and another group of 143 sophomore girls of the same institution with practice teaching experience of from 18 to 36 weeks. They found that there was no significance in the relative scores of the two groups although the reliability coefficient for the test was rather high. They concluded, however, that the test was a better test of mental ability than it was of teaching aptitude.

A second somewhat distinct type of aptitude test available is the Vocational Interest Blank, developed by E. K. Strong, of Stanford University. The test is broken up into eight subtests which attempt to discover interest characteristics in a number of different categories. Three reactions are possible for each situation—like, indifference, or dislike. The subject is supposed to register his first reactions to the situation without permitting time for a studied response. Interest profiles have been established for at least 22 vocations and scoring keys provided, one of which is for teachers.

This test is validated by correlation with characteristic interests of a selected successful group within a given vocation or profession as the criterion. Norms are given with the score key for teachers.

The scale is based on the records of 193 educators, nearly all of whom are members of Phi Delta Kappa (the national professional education fraternity). Approximately one-half are school administrators, one-third teach education in normal schools and colleges, and one-fourth teach in high schools or grammar schools.

The third and first quartiles and median scores are given with grades—steps A, B, B+, B−, and C—established with the per cent of the total distribution of marks given for each grade. Unfortunately no further data were available to the writer on the reliability or validity of the test for teachers. However, Strong in his Manual for Vocational Interest Blank, September, 1930, page 10, makes this cautious statement:

It will be some time before the validity of this test can be exactly determined. Results so far obtained show that the test has genuine merit.

The validity of the test rests finally upon the assumption that there are dominant interest characteristics peculiar to vocational and professional groups, and that these dominant interests are either native or are so early and firmly established in the individual that at least by the time later adolescence is reached these interests have developed fixity or essential permanency. That the permanent nature of these dominant interests is accepted by psychologists in the vocational field is now quite generally recognized. The development of predictive tests based upon basic interest factors in the teaching profession gives promise of significant results.

A third type of predictive test developed for teachers is best represented by Dr. Elizabeth Morris' trait index L test. This test assumes to indicate the presence of certain personal traits considered essential to the successful teacher. As Doctor Morris explains:

The concept of leadership is a useful way of designating or referring to those forms of behavior which include broad interests, control of feeling, tactful management, readiness and ability to undertake activities (often called initiative and resourcefulness), cooperativeness, enthusiasm, sympathy, and the like. Moreover, each of these terms refers to forms of behavior that may contribute to success in teaching (and—in terms of equivalent situations—to success in other professions). Therefore, reactions to a series of situations involving these tendencies as they occur in teaching, are indicative of probable success, especially in that profession.

This point of view is further set forth in these words:

Personality is the total blend of reaction tendencies, and these tendencies must be measured in terms of definite situations. An individuals' personality to a great extent reflects the kind of stimuli which his environment constitutes for him. There is some recognition of this interplay of outer conditions and inner tendencies in explanation of the success of a student under one set of conditions whereas there was failure under different conditions. Selection of teachers, however, will naturally be guided by the standard, "We desire persons who are

most apt to succeed in the various probable teaching situations." It is important, therefore, to measure as many significant tendencies of the individual as possible.

The general quality of leadership was accepted as a complex trait around which centered more specific personal qualities of "resourcefulness (including much of inventiveness and originality), insight, tact, degree of positiveness, and certain emotional attitudes" (13). The trait index is composed of five sections. Section I contains a list of 56 items for which the subject indicates his likes or dislikes, much as in the Strong interest blank, except a range of five rather than three "aspects of feeling" may be recorded for each item. Doctor Morris suggests that this procedure was based on the rather general view, "Tell me what a person likes and I will tell you what kind of a person he is." Section II attempts to evaluate the subject's resourcefulness, insight, and attitudes. Sections III, IV, and V are designed in similar fashion to measure other qualities that enter into the composite trait—leadership; such as in III, tact, initiative; in IV, degree of positiveness of judgment; and in V, characteristic-feeling attitudes.

Doctor Morris found a correlation between the trait index L test and practice teaching grades of 0.463 ± 0.068 (p. 3). While recognizing the weaknesses of practice teaching grades as a criterion, the author justifies her procedure by a quotation from Doctor Whitney's study of 1924:

> Whitney's statistical study shows that grades taken from even geographically scattered schools justify the following comment: "The correlation between teaching success and student teaching remains the highest correlation when all other variables are kept constant."

Although Doctor Morris employed the most approved techniques in this exhaustive study one can not be greatly impressed with the value of trait index L as a predictive device for the selection of aspirants to the teaching profession. In the first place, while validated against the criterion of grades in cadet teaching, the reliability was established with only "60 college seniors preparing to teach in high schools, selected because various kinds of measures were available for each of them" (14). In fairness to the work of Doctor Morris it needs be said that other groups, some much larger in number, were used to determine the value of certain personal traits for use in the test. Secondly, the criterion of cadet teaching success is not a reliable enough standard by which to validate a test which involves reactions to emotional situations. Thirdly, Doctor Morris admits that it was extremely difficult to discover characteristic and sharply accentuated differentiation of response patterns between groups of recognized good and poor teachers. Finally, the low correlation of the test 0.463 does not suggest a predictive device of great value. However, there seems to be much of worth suggested in the theory that lies back of the

construction of this test. Successful teaching does call into play certain personality traits which all are ready to acknowledge though these traits can be defined or identified but vaguely. Further research in this direction may prove most valuable in the development of satisfactory teaching aptitude tests.

An example of the fourth type of teaching aptitude test available is Coxe-Orleans prognosis test of teaching ability. Except that it is labeled a predictive measure of teaching aptitude, a discussion of the test would seem not in place in this paper. In fact, its provision for serious discussion may more properly belong to the paper of the preceding speaker. Back of this test is the implicit assumption that scholastic achievement is a valid criterion of future teaching success. This assumption is definitely set forth by the authors in their Manual of Directions, page 4, as follows:

> It is understood that the value of such predictive measures is greater if they deal directly with the student's ability to teach than with his ability to master the work of the teacher-training institution. As data are obtained of the value of these measures in predicting success in teaching they will be made available. For the time being the contention is offered that the student who does well in the work of the teacher-training institution is more likely to be successful in teaching.

The test is divided into five major parts with several of the parts divided into smaller sections. The plan of the test is essentially that of a survey of the subject's knowledge of facts and theories of education.

Part I. General information test. It presents a point of view and acquaintance with generally accepted opinions relative to education.

Part II. Professional interest test. This also presents a point of view and acquaintance with generally accepted opinions relative to education.

Part III. Lessons in education. Here a series of situations are presented and questions to be answered given.

Part IV. Reading comprehension. Paragraphs are given about which the students' powers of analysis and understanding are tested.

Part V. Problems of education. This test presents crucial problems in American education and then a series of questions secures judgments from the subjects on these issues.

The test requires three hours to complete, is well organized and thorough. It may be valuable as an instrument with which to predict the academic success of students but is of doubtful value as a measure of teaching aptitude. Even as a predictive measure of academic achievement in the normal school the authors present evidence to show that the Terman group test of mental ability is slightly better than the aptitude test as a predictive device of achievement success in the normal school of New York State.

V. GENERAL CONCLUSIONS

1. Interest in predictive factors of teaching success is of comparatively recent date. Very little had been done prior to 1925. Since that time major interest has developed in this field.

2. In some measure the work in this field has paralleled the transitional development of mental tests and aptitude tests in other vocations. Two broad trends have been in evidence in most studies of teaching: (*a*) The attempt to discover teacher traits or qualities directly related to teaching, and (*b*) the attempt to devise tests of a somewhat general nature which would measure or predict teaching success.

3. Thus far the studies of traits and factors predictive of teaching success have not been reassuring. Intelligence, general scholarship, and achievement in professional courses have shown disappointingly low correlations with later teaching success. Practice teaching experience alone has shown a significant relationship to later teaching success. Unfortunately this factor does not have high enough predictive value to be used with confidence and, since it comes at the close of the training period of the prospective teacher is of no apparent value in the selecting of those who should enter the professional training in education.

4. The attempt to devise tests of measurement of general teaching aptitude appears, in the light of the history of mental testing and present research results in this field, to hold greatest promise for the future. As yet the aptitude tests available are at best crude and of little value, though two or three are suggestive.

5. A more serious situation faces the training schools charged with responsibility for educating aspirants for the profession. Adequate curricula can not with confidence be provided until better knowledge is available of those elements of training contributory to success in teaching. Further careful research would appear to be the prerequisite to the solution of the problem.

BIBLIOGRAPHY

(1) Boardman, Charles Willis. Professional Tests as Measures of Teaching Efficiency in High School. New York, Teachers College, Columbia University, 1928. Contributions to Education, No. 327. 85 p.

(2) Boyce, A. C. Methods for Measuring Teachers' Efficiency. Fourteenth Yearbook, National Society for the Study of Education, Part II, 1915. 83 p.

(3) Cahoon, G. P. Marks in College as a Factor in the Prediction of Practice Teaching Success. University High School Journal, May 1930, p. 28.

(4) Engelhart, Max E. Standardized Tests for Students of Education. Educational Administration and Supervision, 15:2:93, February, 1929.

(5) Hamrin, S. A. A Comparative Study of Ratings of Teachers in Training and Teachers in Service. Elementary School Journal, September, 1927, p. 29–44.

(6) KNIGHT, F. B. Qualities Related to Success in Teaching. New York, Teachers College, Columbia University, 1922. Contributions to Education, No. 120, p. 1–4.
(7) ——— p. xiii.
(8) ——— p. 26.
(9) ——— p. 63.
(10) LUDEMAN, W. W. Selection and Elimination of Teacher-Training Material. Educational Administration and Supervision, 13:2:120–124, February, 1927.
(11) MERIAM, J. L. Normal School Education and Efficiency in Teaching. New York, Teachers College, Columbia University, 1906. Contributions to Education, No. 1. 152 p.
(12) ——— p. 53.
(13) MORRIS, E. H. Personal Traits and Success in Teaching. New York, Teachers College, Columbia University, 1929. Contributions to Education, No. 342, p. 13–14.
(14) ——— p. 3.
(15) PYLE, W. H. The Relation Between Intelligence and Teaching Success. Educational Administration and Supervision, 14:4:257–267, April, 1928.
(16) SOMERS, G. T. Pedagogical Prognosis: Predicting the Success of Prospective Teachers. New York, Teachers College, Columbia University, 1923. Contributions to Education, No. 140. 129 p.
(17) ——— p. 125.
(18) ULLMAN, ROY R. The Prognostic Value of Certain Factors Related to Teaching Success. Ashland, Ohio, The A. L. Garber Co., 1931. 133 p.
(19) WHITNEY, F. L. The Prediction of Teaching Success. Journal of Educational Research Monographs, No. 6. Bloomington, Ill.: Public School Publishing Co., 1924. 85 p.
(20) ——— p. 20.
(21) JACOBS, CHARLES L., The Relation of the Teacher's Education to her Effectiveness. New York, Teachers College, Columbia University, 1928, Contributions to Education, No. 277, p. 25.

○